Winter 1977

RECOGNITION LATENCIES, PSYCHOLINGUISTIC REALITY AND STATISTICAL GENERALITY: A PROBE STUDY OF ADVERBIAL CLAUSES WITH LANGUAGE MATERIALS ANALYZED AS A RANDOM FACTOR

JANET MARIE LANG

RECOGNITION LATENCIES, PSYCHOLINGUISTIC REALITY AND STATISTICAL GENERALITY: A PROBE STUDY OF ADVERBIAL CLAUSES WITH LANGUAGE MATERIALS ANALYZED AS A RANDOM FACTOR

by

JANET MARIE LANG
B.A., Boston State College, 1969
M.A., University of New Hampshire, 1974

A THESIS

Submitted to the University of New Hampshire
In Partial Fulfillment of
The Requirements for the Degree of

Doctor of Philosophy
Graduate School
Department of Psychology
December, 1977

302

This thesis has been examined and approved.

John Limber

Thesis director, John Limber, Assoc. Prof. of Psychology

Ronald Shor

Ronald Shor, Prof. of Psychology

G. Alfred Forsyth

G. Alfred Forsyth, Assoc. Prof. of Psychology

Karl Diller

Karl Diller, Assoc. Prof. of English

Merrill Garrett

Merrill Garrett, Assoc. Prof. of Psychology

November 30, 1977

Date

ACKNOWLEDGMENTS

To Jim, simply and completely.

I wish I were a poet--just for these next few minutes. I would write memorable stanzas about the most extraordinary person in the world: Jim Blight, my husband. During all phases of this research, Jim has been my most valuable resource. He humbly denies having any expertise in the area of psycholinguistics, and yet it was during the daily talks with Jim, on our long walks through Grand Rapids, that the ideas presented in this thesis germinated. Jim has been interested in my work--not intrinsically, since he is a historian by nature and choice. His interest in psycholinguistics derives from his love for Janet Lang. I know it, feel it, and I am grateful. Tangible evidence of Jim's influence can easily be found throughout the thesis. I'm glad. He is a part of me, and it should show.

I began this dissertation in Durham, New Hampshire, and finished the writing of it in Grand Rapids, Michigan. I am grateful to the members of my doctoral committee for the interest, advice, and support that they have shown and given to me through this journey.

Dr. John E. Limber of the University of New Hampshire has been a fine chairperson. Calm and knowledgable, he has helped me with large theoretical problems and with nit-picking computer malfunctions. I have benefited greatly from

our association.

Originally, I asked Dr. G. Alfred Forsyth of the University of New Hampshire to be on the Committee to assist me with some of the statistical problems associated with my research. He has done that and much more. From Colorado, New Hampshire, and British Columbia, Al has corresponded with me about all aspects of the research. He has contributed greatly to my education.

Dr. Ronald E. Shor of the University of New Hampshire is a 'Renaissance reader'--he seems to be simultaneously sensitive to gaps in logic, clarity, and grammatical structure. Though such gaps are undoubtedly still present, they are but a small percentage of what was there before Ron's careful reading.

Each of the other committee members, Dr. Karl C. Diller of the University of New Hampshire and Dr. Merrill F. Garrett of The Massachusetts Institute of Technology, has contributed their time and expertise to the project. Each has read drafts and even pre-drafts and on all occasions been helpful and supportive.

Moreover, when I returned to Durham for my final oral, each committee member was extraordinarly generous with his time. Hours of pre-oral meetings helped to make the oral quite enjoyable (at least after the first half hour).

My mother ought to be an honorary member of my doctoral committee. She has been as conscientious as any committee member in gently reminding me of deadlines ("are

you finished yet?"); and she has been outrageously supportive.

Stephanie Bradley-Swift has done the typing. Steph is an expert and a friend. Years ago, she typed Jim's dissertation; for old times sake, she said she had "one more dissertation" left in her. Thanks Steph.

Dan Swift has diligently taken care of all of the administrative details that go along with getting a dissertation approved by the Graduate School. I sat in Grand Rapids, Michigan, and Dan walked all around Durham, New Hampshire. Thanks Danny.

And finally, . . . thank you Jim.

TABLE OF CONTENTS

List of Tables

List of Figures

ABSTRACT

RECOGNITION LATENCIES, PSYCHOLINGUISTIC REALITY AND STATISTICAL GENERALITY: A PROBE STUDY OF ADVERBIAL CLAUSES WITH LANGUAGE MATERIALS ANALYZED AS A RANDOM FACTOR

by

JANET MARIE LANG

A psycholinguistic account of sentence interpretation is constrained by facts about natural languages and facts about mental processing. The effects of a structural variable, clause type (subordinate or main), and a perceptual variable, clause order (first clause or second) in complex sentences with adverbial clauses were investigated using the probe-latency paradigm. Analyses of reaction times yielded a terminative interaction between the two variables. That is, for main clauses only, probes from less recent clauses evoked reliably longer reaction times than probes from the most recent clause; and, only for the most recently heard clauses, subordinate-clause probes evoked reliably longer reaction times than main-clause probes. Probes from clauses that were less recent and subordinate did not evoke reaction times that were different from either most recent subordinate-clause probes or less recent main-clause probes. These

results were discussed in relation to theoretical models of immediate storage and retrieval of sentential material.

The methodological bases for including filler sentences was illustrated. It was suggested that the inclusion and analysis of filler sentences measurably increased both the internal and external validity of the experimental analysis.

Exploratory research regarding the effects of various types of probe words was reported. Probe words that rhyme with words in the sentence had the longest reaction times (as compared with probes either from the sentence or not from the sentence and not related to any word in the sentence). This effect corresponded with a previous finding that rhyme probes were quite difficult for students who were learning English as a second language. It was suggested that the magnitude of this confusion effect may be of use as an index of listening comprehension skills.

In all analyses (experimental and filler), both subjects and language materials were considered to be random or sample variables. $\underline{F}_1$, $\underline{F}_2$, and min $\underline{F}'$ statistics were computed for all effects. In all cases, effects that were significant in one analysis were significant in all analyses; effects that were nonsignificant in one analysis were nonsignificant in all analyses. Results supported the view that if the power of the underlying $\underline{F}_1$ and $\underline{F}_2$ analyses are comparable, then the min $\underline{F}'$ is not an overly stringent test of experimental hypotheses, although it is a mathematically

conservative test.

Statistical and empirical approaches to the problem of generality were critically examined. It was concluded that if prescriptive formulae for generality are to be useful, they must be embedded in explicit theoretical conceptions of the phenomena under investigation.

INTRODUCTION

This dissertation reports and discusses the results of a probe study of complex sentences with adverbial clauses in which the language materials used were analyzed as a random factor. The aims of the study are twofold: to investigate two psycholingustic variables related to sentence interpretation; and to explicitly apply to that investigation some methodological aspects of experimental design and analysis that are related to the problem of generality.

Before the joint aims of the present study can be coordinated, however, one needs to consider a bit of the background that underlies the separate areas of psycholinguistics and research methodology. This introduction is intended to provide such a background.

Part 1--Language and Psychology: A Necessary Merger --presents an overview of the subfield of psycholinguistics. In particular, the dialectic relationship between theories of language and theories of mind is developed. A consequence of this dialectic exchange is that viable theories of sentence interpretation must attend to both the linguistic structures and the perceptual strategies of a listener.

In part 2--Psycholinguistics and the Experimental Method: The Problem of Generality--induction, as classically characterized by Hume, is considered within the context of experimental research. In particular, the inductive leap

involved in applying one's research findings to subjects and materials not sampled is examined. Statistical and empirical approaches aimed at legitimizing that leap are critically reviewed.

Finally, in part 3--The Focus of the Present Study--critical points from theoretical psycholinguistics (part 1) and research methodology (part 2) are coordinated. The probe paradigm is selected as the tool for investigating linguistic structures and perceptual strategies involved in sentence interpretation. Further, the experiment is explicitly designed to include sufficient power for a stringent test of effects over subject and material populations.

Linguistics and Psychology: A Necessary Merger

Psychological interest in language is predicated on the obvious observation that humans engage in verbal dialogues. A speaker has something to say, and says it; a listener hears what was said, and understands it. The listener then becomes a speaker; the speaker a listener, etc. Yet, there is a striking temporal asymmetry in this predication. Verbal dialogues are as old as the human species, psychological interest in such dialogues, however, is a mid-twentieth century phenomenon.[1] The forces behind this recent

[1]See Chapter 2 of Fodor, Bever, and Garrett (1974) for a brief account of psycholinguistics prior to 1960. The territoriality of both linguists and psychologists is obvious: both saw little to be gained from interdisiplinary work. In the 1950's however, there was an overt attempt to combine work in taxonomic grammar with Hullean psychology. Although that specific merger was a failure, it did provide

rapprochement involve a convergence of developments in psychology and linguistics.

The Mind as a Topic of Inquiry for Psychologists and Linguists

During its brief history, experimental psychology has cyclicly embraced the study of mind as central to the discipline and rejected the mind as a construct necessarily outside the realm of scientific inquiry.[2] Over the last three decades the study of mind has once again been recognized as a legitimate topic for psychological inquiry. But modern psychology is not merely an echo of earlier conceptions. Rather, while acknowledging intellectual ancestors,

---

for an integrative framework. The goal of understanding how a speaker-listener makes use of the formal structures of a language continues to characterize psycholinguistics.

[2]Consider the following definitions of psychology:

> In psychology, man looks at himself as it were from within, and he tries to explain the connections among those processes which this internal observation presents to him. (Wundt, 1874, p. 1)

> Psychology is the Science of Mental Life, both of its phenomena and of their conditions. (James, 1890, Vs, p. 1)

> Behaviorism . . . holds that the subject matter of human psychology is the behavior of the human being. Behaviorism claims that consciousness is neither a definite nor a usuable concept. (Watson, 1930, p. 2)

> Psychology is concerned with establishing relations between the behavior of an organism and the forces acting upon it. . . . If I can't give a clean-cut statement of a relationship between behavior and antecedent variables, it is no help to me to speculate about something inside the organism which will fill the gap. (Skinner, from Evans 1968, pp. 21, 22)

theorists have aimed at producing explicit models to characterize the relationship between inner human forces and the outside world. The very title of the watershed book by Miller, Galanter and Pribram (1960), Plans and the Structure of Behavior, suggests that psychologists ought to be concerned with more than just the behavior that organisms exhibit. In their first chapter the authors acknowledge that, by stressing the connection between intentions and behavior, they are in some measure endorsing issues that concerned William James (1890) in his ideo-motor theory. Yet they accurately note that "the bridge James gives us between the idea and the motor is nothing but a hyphen" (p. 12). To explicate that "hyphen," Miller et al. propose a computer-based model (the TOTE) whose input is intentions and output is behavior. In 1967 Ulrich Neisser published a textbook for the field of cognitive psychology. In it he stated that because ". . . the climate of opinion has changed," he needed no ". . . chapter of self-defense against the behaviorist position." Rather, ". . . The basic reason for studying cognitive processes has become as clear as the reason for studying anything else: because they are there" (p. 5).

A similar mentalistic revival has occurred in linguistics. Based largely on the work of Noam Chomsky (1957, 1965), generative grammars have, for the most part, replaced structural grammars. These generative grammars are explicitly concerned with the underlying linguistic intuitions of

humans rather than the classification of observable speech events. Language, generatively described, is seen as a uniquely human mental ability. The creative aspect of language--that humans routinely produce and understand novel utterances--is emphasized. Moreoever, in Chomsky's view, a consequence of embedding language in a mentalistic framework is that the study of language becomes more than the study of isolated words or sentences:

> . . . the study of language . . . will bring to light inherent properties of the human mind . . . Contemporary work in grammar . . . attempts to formulate principles of organization of language which, it is proposed, are universal reflections of properties of mind . . . Viewed in this way, linguistics is simply a part of human psychology: the field that seeks to determine the nature of human mental capabilities and to study how these capacities are put to work (1972, p. 103).

Chomsky's definition of the field of psychology fits well with the conceptions of Miller et al. and Neisser. In fact both books include a chapter on Chomsky's generative grammar and its relation to psychology.

Language: A Part of Cognitive Psychology

The logic of the bond between psychology and linguistics seems to be this: grammar is concerned with those mental abilities that are linguistic; the larger notion of mental abilities subsumes the notion of language abilities; thus, the study of mental abilities (cognitive psychology) naturally includes one of its most important constituents--language (linguistics). Although this relationship is, in general, now obvious and compelling, the details of the psycholinguistic approach often need considerable explication.

What, exactly, do nouns, verbs and dangling participles have to do with psychology? Recently, Fodor (1975) has addressed these issues in a work he describes as "unabashedly an essay in speculative psychology," in which he attempts to discover "how the mind works insofar as answers to that question emerge from recent empirical studies of language and cognition" (p. viii). He argues that there is a kind of logical hierarchy among natural languages, theories of communication, and theories of cognition; details from each preceding level provide empirical constraints on that which follows.

Consider the diagram in Figure 1. Using natural language as a mediator, a message has been transferred from one person to another. This occurs because the utterance conforms to certain conventions of the language shared by both speaker and listener. Linguistic theory is connected with the theory of communication in that generative grammar seeks to explicitly describe conventions or descriptions that must be met by an utterance in order that it be considered part of a language. It is the grammar, then, that attempts to specify the correspondance between message and utterance. The grammar, that is, provides a structural description of each sentence, where "structural description" is defined as a finite set of descriptive levels at which sentences of the language are analyzed. These levels include the quite concrete--those associated with the form of an utterance (e.g., phonetic representations)--and quite abstract levels--those

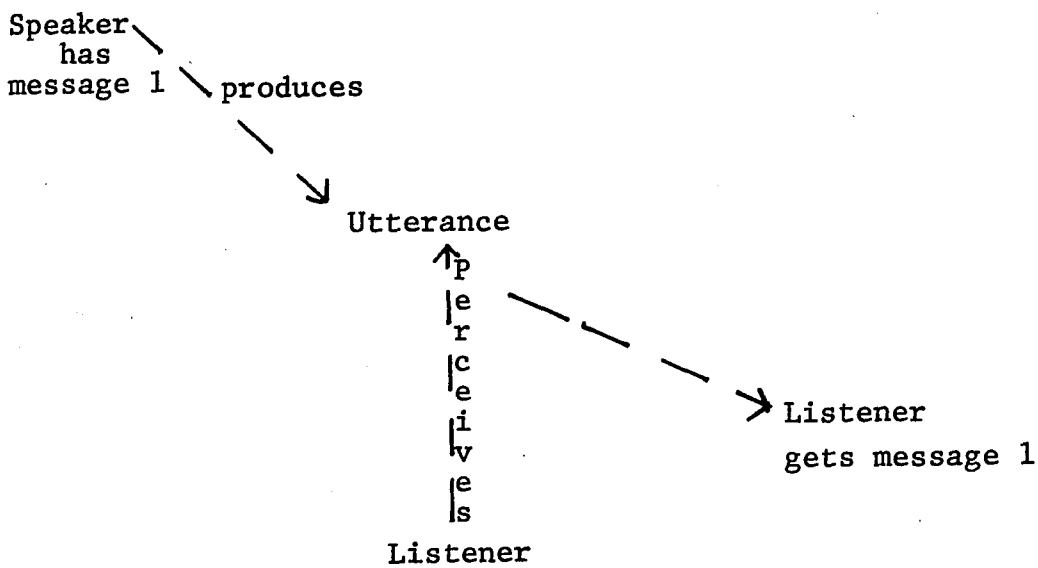


Figure 1. Language as a medium for transmitting messages.

related to the meaning of a message (e.g., deep structure representations).

Because the structural description of a sentence mediates the communication process, it is said to be "psychologically real." It contains, that is, levels which must be involved in processing natural sentences.[3] It is important to note that psychological reality has no necessary relation to awareness. Suppose a speaker utters the word "hat." When I hear that word I intuitively and instantaneously understand that utterance to mean the yellow piece of cloth on my head. Thus, there must be some level within the structural description that represents the utterance as a reference expression. But what of those lower levels that are included in the structural description? One may wonder whether they are really necessary. Is it necessary, that is, in the case of "hat," to postulate a description at the phonetic level such as:

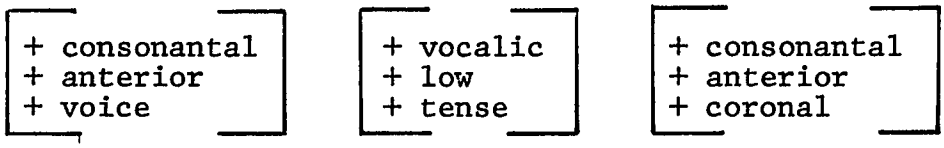
$$\begin{bmatrix} +\text{ low} \\ +\text{ continuant} \end{bmatrix} \quad \begin{bmatrix} +\text{ vocalic} \\ +\text{ low} \\ +\text{ tense} \end{bmatrix} \quad \begin{bmatrix} +\text{ consonantal} \\ +\text{ anterior} \\ +\text{ coronal} \end{bmatrix}$$

It is obvious that the utterance "hat" may be described as a series of specific sounds. In fact, the proper utterance requires that a speaker meet certain phonetic demands. Yet, the details of the phonetic requirements are of little

---

[3]See Fromkin and Rodman (1974) for a discussion of the psychological reality of distinctive features, phonemes, and syllables. Subsequent sections in the introduction examine the psychological reality of specific syntactic units.

interest to either the speaker intending to convey some information or to the listener attending to the message. Suppose, however, that the message misses the phonetic mark, as in the following sequence:



In this case, the reference is misinterpreted. Thus, although not intending to utter a sequence that satisfies certain phonetic requirements, a speaker must do so if referential messages are to be accurately transmitted. The lower concrete descriptions thus become the necessary means to the intended, abstract, referential end.

Expansion of an utterance into a phrase or a clause also requires the postulation of intermediate, syntactic levels of description. A speaker may not intend, for instance, to convey information about the tense of a verb, but unless the utterance includes such syntactic information, normal ease of communication will be interrupted. Consider the difference beteen the following sentences:

1a. It is a pleasure to write.

1b. It is a pleasure to have written.

At this moment I would deny the first message but heartily endorse the second. If I am to communicate these ideas to a listener, my utterances must satisfy certain syntactic requirements relevant to the tense of the verb. Moreover, the

listener must recognize that the utterances do so.

It is important to note that psychological reality is claimed for levels of a structural description but not necessarily for operations that connect the levels with each other.[4] Thus, a grammar, bound by the constraints of natural language, is subsumed by a broader theory of communication. That theory, in turn, is bound by the constraints of messages and it is subsumed by a broader theory of cognition. Messages are ideas--part of the mind--and the means by which they are represented is a mental process. Theories of mental processes, therefore, must account for the means by which messages are interpreted, and it is assumed that information about linguistic abilities provides important boundaries for cognitive theories.

One of the things we can do with linguistic material is forget it. But the forgetting is not random and it is

---

[4]In nearly everyone of his books, Chomsky has emphasized that a grammar is not intended to be a model of a speaker/hearer. The following quote is representative:

> To avoid what has been a continuing misunderstanding, it is perhaps worth while to reiterate that a generative grammar is not a model for a speaker or a hearer. It attempts to characterize in the most neutral possible terms the knowledge of the language that provides the basis for actual use of language by a speaker-hearer. When we speak of a grammar as generating a sentence with a certain structural description, we mean simply that the grammar assigns this structural description to the sentence. When we say that a sentence has a certain derivation with respect to a particular generative grammar, we say nothing about how the speaker or hearer might proceed, in some practical or efficient way, to construct such a derivation. (1965, p. 9)

related to the abstract levels of the structural description of the message. (I can't remember many specific words or sentences, but I remember reading last night that the Red Sox lost to Mark Fidrych and the Tigers!) It is probable, therefore, that the capacity to forget (selectively) is related to the way in which information is stored.[5] The following diagram (Figure 2) is meant to convey schematically the interdependence of the signal, grammar, and memory. It is obvious that the study of language is relevant to the study of mind. Whatever else they may do, theories of cognition must account for the manner in which ordinary people utter and interpret natural language.

Cognition: A Part of the Psychology of Language

Thus far, the models of the speakers and listeners have included only their linguistic knowledge. Bever (1970) has suggested that there is much more to the ordinary use of language. He has argued for the "cognitive basis for linguistic structures" (p. 279), suggesting that the relationship between language and mind is dialectical: linguistic information constrains cognitive theories and perceptual/cognitive information constrains linguistic theories.[6] Even

---

[5]See Fodor et al. (1974), especially the section entitled "The coding hypothesis" (pp. 264-268).

[6]This is not to imply that Fodor (1975) neglects the language and mind dialectic. Support for the interplay is implicit throughout the book. Rather, the point is that Fodor emphasizes that linguistic information constrains cognitive theories; and Bever emphasizes that cognitive information constrains theories of language. The two together constitute a dialectical approach.

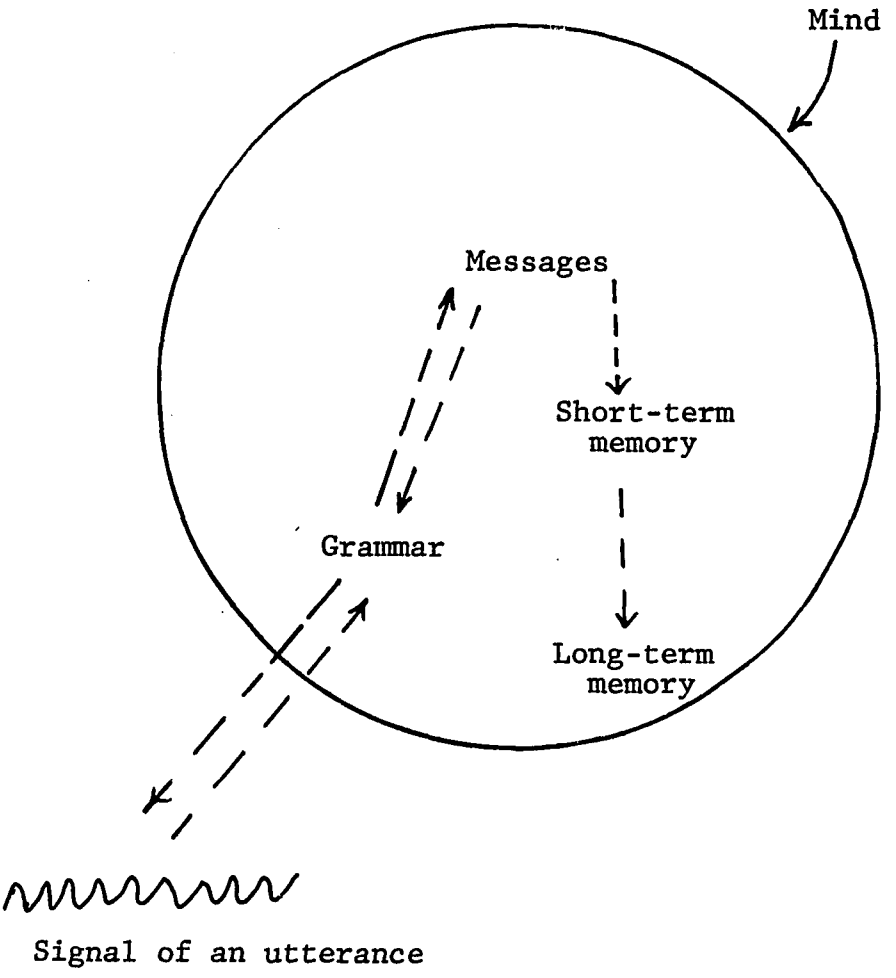


Figure 2. The interpretation of an utterance.

the formal definition of language, in Bever's view, cannot be isolated from other cognitive skills. Any attempt to specify pure linguistic structure is therefore artificial and problematic:

> Certain ostensibly grammatical structures may develop out of other behavioral systems rather than being inherent in grammar. That is, linguistic structure is itself partially determined by the learning and behavioral processes that are involved in acquiring and implementing that structure (1970, p. 280).

Thus, a model of a speaker/listener must include not only intuitive linguistic knowledge (the domain of a grammar), but also a means for the implementation of that knowledge (the domain of perception/cognition).[7] An examination of the psychological reality of linguistic structures must be integrated with a consideration of the psychological reality of linguistic processes. One important goal of such an integrated approach is to eliminate descriptions of processes that are "technically accurate but perceptually irrelevant" (Bever, 1973, p. 150).[8]

---

[7]Fodor et al. (1974) suggest that this psycholinguistic approach is a useful and prototypic psychological approach to non-linguistic areas as well:

> . . . any serious account of the behavior of an organism will have to say not only how (i.e., by virtue of what psychological mechanisms) the organism puts its knowledge to use, but also what it is that the organism knows; what concept it has (p. 5).

[8]Bever used the quoted expression to describe a non-Gestalt or elemental approach to perception. I believe that by analogy the expression is also applicable to a non-perceptual approach to sentence interpretation.

The Unreality of a Purely Linguistic Account of Sentence Interpretation. The initial phase of contemporary psycholinguistic inquiry was anything but integrated; the early work failed to heed Chomsky's warning that generative, transformational grammar is not a model of the way in which speakers and listeners produce and understand sentences (see footnote 4). Some psychologists, impressed with the apparent power and generality of his theory, sought to test the possibility that transformational complexity is isomorphic with perceptual complexity. Research on this hypothesis, the Derivational Theory of Complexity (DTC) constituted this first phase of post-Chomskian psycholinguistic inquiry. Sentence pairs exhibiting various syntactic constructions were examined. In each case, the application of at least one extra transformational rule was necessary to derive one pair member. The DTC always predicted that the sentence with the longer derivational history would be perceptually more complex. The comparison of active and passive sentences was prototypic; passives, it was predicted, would be harder to understand than actives. Although there was some initial empirical support for this position (Miller & McKeen, 1964), researchers soon found that the perceptual asymmetry between actives and passives holds only in special cases (Slobin, 1966). In particular, sentences were seen as falling into one of two categories--reversible and irreversible. A reversible sentence is one in which an equally plausible sentence would result if the subject and object phrases were

interchanged. For example:

2a. The tall man saw the girl with curly hair.

2b. The girl with curly hair saw the tall man.

2c. The girl with the curly hair was seen by the tall man.

2d. The tall man was seen by the girl with the curly hair.

In an irreversible sentence, an exchange of subject and object phrases results in an implausible sentence. For example:

3a. The tall man climbed the fence.

3b. The fence climbed the tall man.

3c. The fence was climbed by the tall man.

3d. The tall man was climbed by the fence.

In reversible sentences, passives are harder than their corresponding actives; in irreversible sentences, there is no difference between actives and passives. Thus, the perceptual difference between actives and passives cannot be due solely to transformational complexity because sentences which do not show this asymmetry are equally complex. One must postulate, it seems, a source of variability other than grammatical rules.

Presumably, in irreversible sentences, the intended message can be inferred directly from the words of the sentence, thus rendering syntactic processing superfluous. In reversible sentences, however, the information provided by the words is ambiguous; an analysis of the syntactic

features of the sentences is necessary.[9] It is obvious, therefore, that predictions based on the DTC are not uniformly confirmed. The passive transformation provides at least one important exception; in it, some non-syntactic, heuristic procedure must be involved. A number of studies (Fodor & Garrett, 1966; Bever, 1970; and Fodor et al., 1974) have found the DTC inadequate in many other respects. They have concluded that whatever the relationship between grammatical complexity and perceptual complexity, it is not as direct as the DTC predicts.[10]

In retrospect, the failure of the DTC is not surprising. Generative grammar was not intended to be a model of the speaker/listener; it is an explicit description of the nature of linguistic structure. Further, the DTC predicted many events which simply are counter to common sense and experience.[11] It must predict, for example, that 4a is

---

[9]Even with reversible actives and passives, it is dubious that it is the passive transformation per se that is responsible for the added complexity. Limber (1977) points out that the "by" in the passive sentence is potentially ambiguous. "By" may indicate either a passive or a locative construction.

[10]History seems to be repeating itself. Within semantics, researchers are investigating the relationship between the surface vocabulary of a language and the semantic representation that the grammar would supply. Fodor (1975) reports: "the predicted correspondence between definitional and perceptual complexity does not seem to hold" (p. 147).

[11]Some of the empirical inadequacies of the DTC have just been described. Limber (1976) considers the logical inadequacies of DTC.

perceptually more complex than 4b, and that 5a and 5b do not differ in perceptual complexity:

4a. Your gold medallion has a lovely inscription on it.

4b. Your medallion which is gold has a lovely inscription on it.

5a. The man the woman the mother liked loved walked.

5b. The race the man the reporter interviewed planned was postponed.

Moreover, the grammar of a language (as conceived by Chomsky, 1947, 1965) is, by design, not equipped to comment on two common linguistic events. (1) The grammar generates some sentences that are perfectly grammatical yet completely uninterpretable by native speakers of the language. Consider, for example, a sentence like 5a but with thirteen embeddings. Formal restrictions in the grammar do not set an upper limit on the number of embeddings per sentence. (2) Certain lexical strings not generated by the grammar are interpretable by native speakers of the language. An example of such a string appears in 6.

6*. The man tall he had on yellow shoes.

Sentences may thus be classified in the following way: grammatical-ungrammatical, acceptable-unacceptable and interpretable-uninterpretable. These dimensions are at least partially independent of one another. The grammar may determine that which a listener extracts from an utterance, but it cannot provide a description of the manner in which the

listener obtains such information. Any comprehensive theory of sentence processing must, therefore, address itself to both grammatical structure and psychological processes--it must be truly psycholinguistical.

Perceptual Strategies and Sentence Interpretation. Bever (1970), Fodor, Bever and Garrett (1974) and Limber (1976) argue for a broad theory of sentence interpretation that considers the interaction of cognitive systems with linguistic information. Sometimes the interactive nature of these variables permits a listener to use non-linguistic information to solve an essentially linguistic problem--as in the clarification of some ambiguous sentences. At other times, input from cognitive systems may reduce the importance of syntactical analyses--as with irreversible passive sentences (see sentence 3c).

It is obvious that in many instances, cognitive and linguistic information are not easily distinguishable. Irreversible passive sentences are understood as easily as their corresponding active sentences because semantic properties of the lexical items limit the plausible actor-action-object relationships into which those words can enter. That is, with respect to sentence 3c, "climbing" is something that a man does to a fence, and not vice versa. A critical question is this: are semantic constraints (such as represented in 3c) examples of the dominance of linguistic information provided by the grammar? Or, conversely, are they

examples of pragmatic information that reflects something about objects and actions in the world? One may assume that pragmatics are always involved--sentence interpretation occurs, in most instances, in the real world of objects and actions. But one must also consider the extent to which a grammar can be explicit about a listener's expectancies. For example, the lexicon entry for the verb "climb" might include a description which specifies the requirement of an inanimate object. In this case, the grammar would be isomorphic with the expectations of the listener and it would therefore disallow sentences 3b and 3d. But now let us consider sentences that contain non-syntactic implausibilities. For example:

7a. The mosquito bit the woman.

7b. The woman bit the mosquito.

As 7b describes a statistically infrequent act, 7a would be labeled irreversible. The grounds for irreversibility are clearly pragmatic; 7b violates an expectation (one may assume) rather than a formal property described by the grammar.

Any comprehensive theory of sentence interpretation, we have seen, must allow a place for linguistic structure and for memory--for one's understanding and prediction of worldy events. The latter are commonly incorporated into the theory as heuristics.[12] A heuristic is a rule of probability based

[12]Bever (1970) uses the phrase "behavioral induction," and admits that the source of the induction (linguistic or

on the expectation that sentences will follow the same characteristic patterns which described them in the past; that is, certain grammatical features are weighted more heavily than others. Since a grammar contains no weighting procedure, the assignment of probabilities must be accounted for by a model of sentence perception. Heuristic procedures discussed by Bever (1970) and Fodor, Bever and Garrett (1974) are associated with three aspects of sentence interpretation: clausal analysis, lexical analysis, and surface structure analysis.

Clausal Analysis: The understanding of an utterance implies that the acoustic signal has been segmented into chunks which correspond to deep structure sentoids.[13] A native speaker can easily delineate the three segments of the following utterance:

8. you are teaching a class I am writing in the library it is hot outside today.

Grammatical knowledge underlies the segmentation--the task could not be performed (at least by me) with a German utterance. The grammar, however, specifies many levels of linguistic segmentation--phonetic, words, phrases, as well

---

experimental) is not obvious. In this regard, consider the following remark from C. S. Pierce, as quoted in Chomsky (1972, p. 91):

> induction has no originality in it, but only tests a suggestion already made.

[13] A sentoid is like a clause. Technically, it is a phrase structure tree that is immediately dominated by a S node.

as clauses. None of these levels is necessarily dominant. Yet experimental and anecdotal evidence indicates that there is a perceptual strategy for segmenting utterances into clauses. This strategy, or heuristic, applies to a discourse which contains a number of sentences with only a single clause (as in sentence 8), and to sentences that contain more than one clause.

Using the "click" paradigm developed by Ladefoged and Broadbent (1960), Garret, Bever and Fodor (1966) were able to show that the processing of a sentence requires that listeners be sensitive to the clausal structure even when the natural prosodic features of the utterance are suppressed. Pairs of sentences with common lexical items but different clause boundaries were constructed:

9a. (In her hope of marrying) Anna was surely impractical.

9b. (Your hope of marrying Anna) was surely impractical.

The common items were cross-recorded so that the acoustic properties of the sentence pairs was identical. An instantaneous burst of noise (a "click") was then placed within the word "Anna." The listener's task was to indicate the location of the click. The rationale is this: if the clause is a psychologically real perceptual unit, then it should be processed as a whole and resist the interruption of the click. Results support this prediction: listener's receiving sentence 9a prepose the click into the clause boundary between

"marrying" and "Anna," while listener's receiving 9b postpose the click into the clause boundary between "Anna" and "was." Since both groups of listeners heard an identical recording, their differential performance must be due to some more abstract dimension along which the sentences differ. The sentences have different clause boundaries and it is likely that, at some level, the listeners know this. Studies such as these suggest that the clause as a structural description is psychologically real. It may be concluded therefore that listeners engage in some sort of clausal grouping.

The "click" studies, important and striking though they are, shed no light on the process that underlies clausal analysis.[14] Bever (1970) has suggested that the segmentation of clauses is accomplished by the following heuristic:

> A Segment together any sequence X . . . Y, in which the members could be related by primary internal structural relations, "actor action object . . . modifier (p. 290).

Various applications of this strategy might account for the segmentation of a single clause sentence or a sentence with more than one clause. Fodor et al. (1974) refer to this strategy as the "canonical-sentoid" strategy. It is simply true that a surface ordering of noun phrase, verb, optional noun phrase usually corresponds to subject, verb, and optional object of deep structure which, in turn, corresponds to actor, action, and optional object of the message. Thus, it

---

[14] Garrett, Bever and Fodor (1966) has been presented as a prototypic click study. For a more extensive review of research using the click paradigm, see Fodor et al. (1974), pp. 329-341.

seems plausible simply on rational grounds that the canonical hypothesis is valid. Moreover, Bever (1968) found that if a canonical trap was set for listeners, their errors, made when paraphrasing a sentence, could be made predictable and more numerous. Specifically, he compared performance on sentence pairs like 10a and 10b.

10a. The editor the authors the newspapers hired liked died.

10b. The editor authors the newspaper hired liked died.

Both sentences have the following syntactic arrangement:

NP NP NP VP VP VP

In 10b, however, the NP "authors" is lexically ambiguous. It could be a plural noun or it could be a third-person singular, present tense verb. Thus, if the sentence is processed sequentially from beginning to end, at least through the middle of sentence 10b, a listener has two syntactic descriptions of the sentence:

NP NP NP . . .

NP VP NP .

The second one conforms to the canonical strategy and should be preferred. Once the last part of the sentence is heard, the canonical hypothesis becomes untenable, and only the NP NP NP description makes any sense. The point is not simply

that 10b is perceptually more complex than 10a. The difference in perceptual complexity can easily be accounted for by the lexical ambiguity of the phrase "authors" in 10b. More importantly, the ambiguity allows the listener to employ the canonical strategy. In addition, that strategy, once used, tends to persist. Listeners mistakenly paraphrase the utterance as though "the editor authors the newspapers" is a complete sentoid, even though this interpretation renders the remainder of the sentence nonsensical. Its resistance to correction indicates that the strategy is fixed firmly in a listener's repetoire of techniques used in the interpretation of sentences.

Of course, the canonical hypothesis is not the only heuristic used; noncanonical sentences are understood. Moreover, as Fodor, Bever and Garrett (1974) point out, listeners do not always fall into canonical traps. The following sentences appear to be approximately equal in complexity:

11a. They wanted John killed.

11b. They wanted John dead.[15]

Yet for one sentence the canonical strategy yields a correct description of subject-verb relations while for the other sentence it does not. If the strategy is applied twice to 11a, the following subject-verb sequences emerge: (they wanted), (John killed). But John is the object rather than the subject of "killed." In 11b "dead" is not interpreted as

---

[15]These sentences are from Fodor et al. (1974), p. 348.

a verb phase, thus "John" is not misinterpreted as the preceding subject NP. A heuristic, then, does not work every time but it works often enough to be useful. A model of sentence interpretation must, therefore, specify the structural organization on which a heuristic operates and it must describe procedures for handling exceptions.[16]

Lexical Analysis: A clausal analysis partitions a sentence into units of a message or proposition. The internal, structural relationship among the elements of each unit must then be determined. In some cases, a knowledge of word meanings and judgments of plausibility combine to produce correct inferences about the grammatical relationships within a clause. For example, the words "man," "climbed," and "fence" are internally related as actor-action-object. The relationship is easily inferred on semantic grounds for both the active and passive forms (sentences 3a and 3c). For the case of irreversible sentences, Bever (1970) suggests "that the presence of unique semantic constraints allows syntactic factors to be bypassed entirely" (p. 296).[17] Bever further speculates that reliance on heuristics is probably much greater in natural conversations than it is with isolated sentences in a laboratory setting; natural conversations provide contextual information and knowledge about the speaker,

---

[16]See Limber (1977) for an example of heuristics that operated on relative clauses.

[17]Bever (1970) also points out that semantic constraints can ease the understanding of some center-embedded sentences. Sentences 5a and 5b (p. 17 of this introduction) illustrate this point.

etc. which is unavailable in the controlled laboratory.

We may wish to conclude from Bever's analysis that, for certain sentences in certain situations, the interpretation of a message occurs without any kind of syntactic analysis.[18] Yet Bever cannot be quite right. Consider the following utterances from which a listener obviously receives an identical message:

12a. The man climbed the fence.

12b. The fence was climbed by the man.

12c. man . . . climbed . . . fence

The listener also, and just as obviously, knows there is a difference in form between these sentences. 12a and 12b are well-formed sentences while 12c is not. Moreover, 12a and 12b contain syntactic noise which is not present in 12c; one might guess, therefore, that the non-noise condition is the preferred form. Yet most individuals do not speak in telegrammese; there is no reason to suppose they understand in telegrammese either. We might tentatively conclude, therefore, that the interpretation of an utterance always involves a device or process which monitors the syntactic correctness of an utterance.[19]

---

[18]Bever is quite clear on this point: "Thus, most normal perceptual processing of sentences is probably carried out with little regard to actual sequence or structure" (1970, p. 297).

[19]Even if Bever's point is legitimate, that a syntactic component need not be postulated to explain the perception of some utterances, there is still the necessity of including such a component in a production model. But presumably, production-model components are present to facilitate perception.

A lexicon contains more than just the dictionary meanings of words. It also includes the syntactic features of each word--a list of the kinds of constructions in which that word can appropriately appear.[20] For verbs, this includes information about the range of possible noun phrases that verbs can take as objects. Consider the following verbs: contain and require. Both verbs are transitive; they take nouns as direct objects, as in 13 and 14:

13. Your paper contains many interesting ideas.

14. My body requires food.

"Require," but not "contain," can also take a compliment construction as an object, as in 15:

15. The instructor required students to write intelligible English sentences.

In this case, the object NP is not just a noun; it is a clause in which [students-write-sentences] form a [subject-verb-object] relationship. We may summarize as follows:

16a. Subject verb (contain) direct object

16b. Subject verb (require) direct object
clause

"Contain" provides explicit structural information about what is to follow; "require" provided ambiguous information.

What is the effect of lexical differences on the processing of sentences? Fodor, Garrett and Bever (1968) found that sentences with complex verbs (i.e., verbs that

---

[20] See J. J. Katz "The Realm of Meaning" in G. A. Miller (ed.), Communication, language and meaning, for an introduction to the use of syntactic and semantic features in a lexiom.

allow more than one type of grammatical construction) are perceptually more difficult than identical sentences with simpler verbs. For instance, 17a which follows is more often erroneously paraphrased than 17b. In addition, when subjects were asked to arrange grammatically words which were presented in random order, they more often failed and started falsely with a-type sentences.

17a. The old theory obviously required several false assumptions about cosmology.

17b. The old theory obviously contained several false assumptions about cosmology.

Note that both sentences have identical syntactic structure. The complex verb, although capable of taking a complement, does not. Yet for most of Fodor, et al.'s sentences, the initial part of the sentence does not exclude the possibility that a complement will follow the verb, as it does in 17c:

17c. The old theory required that somewhere there should be peanuts without shells.

It is not so much the complexity of the verb, per se, therefore, that increases the perceptual load; rather, the lack of cues early in a sentence inhibits a listener's capacity to anticipate the proper construction. Indeed, once a complex verb has been identified (or any lexical item that is compatible with more than one construction) by a listener, one may reasonably ask how that person ever arrives at a correct analysis of the sentence. Surface structure cues can be very helpful here.

Surface Structure Analysis: A listener's task is to derive a message from an uttered speech signal. The utterance must be segmented into clauses, structural relations must be assigned to elements in each clause and, in complex sentences, the relationship between the clauses must be considered. Information necessary to perform these tasks must be present in the surface structure and the listener must be able to extract and use that information. Surface structure cues may be relatively obvious or subtle; as they become more obvious they render a sentence less difficult perceptually.

A verb can be used in a variety of deep structure constructions. How, then, is a listener to know which construction is appropriate for the analysis of any particular sentence? Recall that the verb "require" can take either a direct object or complement objects (to- of that- complements). In sentence 18a which follows, the word "to" clearly marks the infinitival construction:

18a. The instructor required the students to work.

Similarly, in 17c, the word "that" signaled a complement construction. Note, however, that in 17c "that" could be deleted from the sentence. Thus, it ought not be the only surface-structure cue signaling the complement clause. Compare 17c (with the word "that" deleted) with 17a. For the simple direct-object construction (17a), there is no verb following the complex verb "require"; but in 17c an additional verb signals another cluase, and hence another construction.

Compare the two versions of 17c (with and without the pronoun, "that"). Both sentences contain surface-structure cues which indicate that the sentence has two clauses. One sentence has two cues, and the additional cue (the "that") occurs at the beginning of the second clause. A theory of sentence interpretation which stresses importance of surface-structure cues would predict that deleting the pronoun "that" from the sentence should increase the perceptual complexity of the sentence. Fodor and Garrett (1967) constructed pairs of center-embedded sentences in which the relative pronouns either were or were not deleted (as in 19a, 19b).

19a. The window which the ball that the boy threw hit broke.

19b. The window the ball the boy threw hit broke.

They found that the presence of the relative pronouns facilitated the speed and accuracy with which subjects paraphrased the sentences. They also compared two versions of 19b--one version read in a monotone, the other read with normal intonation. They found that the intonation (another surface-structure cue) facilitated the comprehension of a sentence, although its effect was inferior to that of relative pronouns in a sentence without normal intonation.

Relative pronouns are powerful surface-structure cues which signal the presence of an additional clause. They also yield information about the relation between or among the clauses in a sentence. Specifically, relative pronouns

introduce subordinate clauses; i.e., those clauses which are always embedded within a main clause. The main clause conveys the primary content of the propositon; the subordinate clause adds supplementary information about the primary content. Subordinate clauses can also be introduced by adverbs, such as:

20a. Before their nine-game losing streak, the Red Sox were in first place.

20b. After we saw "Fiddler on the Roof," we began reading World of Our Fathers.

In these sentences, the initial adverb clearly marks the subordinate-main structure. Notice that 20a and 20b could be reversed, making the clause order main-subordinate.

20c. The Red Sox were in first place, before their nine-game losing streak.

20d. We began reading World of Our Fathers, after we saw "Fiddler on the Roof."

In these cases, if the only heuristic a listener uses is to look for the presence of an adverb or a relative pronoun that introduces the subordinate clause, then for 20c and 20d no decision about the sentence can be made until the mid-point of the sentence.

A heuristic has been suggested by Fodor, et al. (1974) which determines the relation between clauses in a sentence, and which account for the cases illustrated in 20a-d:

> B Take the verb which immediately follows the initial noun of a sentence as the main verb unless there is a surface structure mark of an embedding (p. 356)[21]

This heuristic is a reflection of both perceptual and linguistic factors. Weksel and Bever (1966) found that for various kinds of complex sentences, main-subordinate orderings (as in 20c and 20d) are preferred over subordinate-main orderings (as in 20a and 20b). Clark and Clark (1968) found that subordinate-main versions were harder to memorize than main-subordinate versions. A listener's working assumption seems to be that the main clause will be presented first, while the subordinate one will come later. If this order is to be violated it must be clearly marked. The syntactic rules of English reflect this working hypothesis:

> A subordinate clause is marked as subordinate by the end of its verb phrase if it is the first clause in a sentence, but may go unmarked if it follows the main verb (Fodor, et al., 1974, p. 358).[22]

That is, the grammar will block the deletion of words that introduce a subordinate clause if that clause is the first clause of the sentence and if, without that introductory marker, the subordinate clause could be mistaken for the main clause. But the deletion will be permitted if the subordinate clause follows the main clause. Thus, in sentences

---

[21]Bever (1970) presents a similar heuristic: "The first N . . . V . . . (N) . . . clause . . . is the main clause, unless the verb is marked as subordinate" (p. 294).

[22]Bever (1970) argues that this syntactic rule is a consequence of the perceptual strategy presented in heuristic B. Determining such temporal order is at least problematic: a chicken versus the egg dilemma.

21a and 21b where the first verb is the main verb, the word "that" is optional; but, in sentences 21c and 21d where the subordinate verb is first, the "that" must remain in the sentence to mark the subordination. Sentence 21d, which would present a problem for heuristic B, is not allowed by the grammar.

21a. Two Red Sox fans in Michigan believed that Boston could still win the division title.

21b. Two Red Sox fans in Michigan believed Boston could still win the division title.

21c. That Boston could still win the division title was believed by two Red Sox fans in Michigan.

*21d. Boston could still win the division title was believed by two Red Sox fans in Michigan.

To summarize, theories of sentence interpretation have advanced quite a bit from the early suggestion that when listeners understand a sentence they are performing computations analogous to the transformational operations that mediate surface and deep structures of a sentence. Rather, the distinction between grammaticality and acceptability has emphasized that at least part of what the grammar generates is beyond the interpretative ability of a listener. Perceptual and cognitive limitations need to be an integral part of the model. Information from past experience, both linguistic and nonlinguistic, needs to be incorporated into the model so that the listener can draw from those sources and make inferences about new linguistic material. Early

psycholinguistic work stressed the importance of the grammar primarily because the contribution of grammar had been previously ignored. Contemporary psycholinguistic work stresses the non-exclusivity of the grammar in a broader theory of communication. Heuristics that deal with clausal, lexical, and surface-structure analyses draw on grammatical and inferential information that the listener has and are readily used in contemporary sentence-recognition models.

Psycholinguistics and the Experimental Method: The Problem of Generality

Ordinarily, any particular piece of research is of interest because it is embedded in some theory. That is, a theory generates some specific predictions about a general class of events, and a research design is constructed to test the predictions over a subset of those events. Experimental results then either support or refute the theory. A vast body of literature on experimental methodology suggests, however, that the presumed direct and automatic link between experimental results and theoretical conclusions is often tenuous (Meehl, 1967).

One recurrent theme deals with the issues of generalizability. To what extent can experimental findings from a relatively small sample reflect relationships in a larger population? The problem centers around the issue of inductive inference, predicting unobserved relationships from knowledge

about observed relationships.[23] Since such predictions necessarily involve extrapolation beyond actual experimental findings, they would seem to require some kind of justification.

Statistical Solutions?

Within experimental psychology, the justification has often taken the form of reliance upon statistical procedures embedded within the design model. In fact, in designing analysis of variance procedures, Fisher explicitly aimed at making induction automatic:

> That such a process induction existed and was possible to normal minds, has been understood for centuries; it is only with the recent development of statistical science that an analytic account can now be given (1955, p. 74).

The analysis of variance is based on the assumption that any experiment is simply one sample from a hypothetical population of experiments. Statistical procedures are then

---

[23] Hume examines the problem of induction more closely, and concludes that even among observable events, the statement of a cause and effect relationship involves a mental construction.

> When any natural object or event is presented, it is impossible for us, by any sagacity or penetration, to discover, or even conjecture, without experience, what event will result from it . . . Even after one instance or experiment where we have observed a particular event to follow upon another, we are not entitled to form a general rule; . . . it being . . . an unpardonable temerity to judge of the whole course of nature from one single experiment, however accurate or certain. But when one particular species of event has always, in all instances, been conjoined with another, we make no longer any scruple of fortelling one upon the appearance of the other . . . We then call the one object, Cause; the other, Effect. We suppose that there is some connection between them . . . this connection . . . we feel in the mind (II, vii).

used to estimate the extent to which that actual experiment is representative of other experiments in the population. Thus, if one meets the assumptions of the model, one is reasonably sure about not only the cause and effect relationships in the experiment, but also about the likelihood of future replications.

The central assumption in Fisher's model, and the one that most researchers are keenly aware of, is that of a randomly sampled subject pool. If subjects are randomly selected and assigned to experimental groups, then one can be confident of two things: 1) that the sample and the population differ mainly in size, and so, a description of one is a description of the other, and 2) that the various treatment groups are relatively equal before any treatment has been introduced. Then, by treating subjects as a random variable, one minimizes the extent to which the results are dependent on the particular subjects used. One is legitimately (if not logically) able to generalize the results to a wider range of people than were tested.

Typically, only subjects are deliberately sampled and analyzed as a random variable. Yet obviously the researcher wishes to generalize across more than just the subject pool. Variables that are thought to be theoretically irrelevant (such as the time, date, and location of the experiment, as well as the experimenter) are assumed to be randomly distributed throughout the population of experiments, and as such

do not critically affect the results.[24] Theoretically relevant variables that are sampled (i.e., the independent variables) are usually analyzed as fixed variables. That is, they are a fixed part of both the actual experiment and all experiments in the hypothetical population of experiments. Thus, when results are obtained, the model predicts that those fixed effects will remain stable in spite of changes in any of the random variables.

Clark (1973, 1976) has argued that in psycholinguistic research, a second random variable ought to be routinely added to the analysis of variance model.[25] Language materials and stimulus materials throughout experimental psychology are selected from a large (perhaps infinite) population and researchers typically discuss results as if the effects are constant across a set of materials larger than those sampled. Thus, the materials variable would seem conceptually like the subjects variable and so should be similarly analyzed. Otherwise, the inductive leap to materials has no statistical legitimacy. Clark proposes an alternative analysis--essentially selecting a more appropriate denominator for the

---

[24]For an alternative account of the role of an experimenter, see McGuigan (1963).

[25]Actually, Clark (1973) is, at least in part, reiterating a methodological issue raised earlier by Coleman (1964). But, although both papers are contentually similar, Coleman's paper scarcely caused a ripple while Clark's paper has stirred up quite a storm. It is Clark's paper that is referred to by both proponents and opponents in the language-as-fixed-factor controversy. This is due to innovative and controversial statistical procedures that Clark has proposed; in part, perhaps also, to Clark's position on the editorial board of the Journal of Verbal Learning and Verbal Behavior.

F ratio used to test for treatment effects. By advocating a statistical remedy for the conceptual and philosophical problem of induction, Clark is echoing Fisher's own belief that the strict adherence to the analysis-of-variance model will result in ". . . perfectly rigorous and unequivocal inference" (1960, p. 4).[26]

If the idea behind Clark's suggestion is rather simple (i.e., substituting one MS for another), the implementation of that suggestion is quite complicated. When a design has two separate random factors (as would be necessary to simultaneously generalize across subjects and materials), there is no single MS that can be used to appropriately evaluate the treatment effect unless the two were perfectly confounded. Winer (1971) suggests the use of a quasi F test (F') where the appropriate MS is estimated by pooling available MS's. But this design assumes a complete data matrix (that there is an entry for each subject-language item combination). As Clark points out, studies using a reaction time measure as the dependent variable often have a number of missing data cells. In these cases, F' can be approximated in a relatively straightforward manner by computing min F'. Min F' is based on the results of two simpler analyses. The traditional analysis with subjects as the sampling variable (and materials as a fixed factor) yields an F_1; a similar

---

[26]Bakan (1966) questions the appropriateness of relying on a statistical test as a means for making an inductive inference. Though his remarks were directed against Fisher, they are similarly applicable to Clark.

analysis with materials as the sampling variable (and subjects as a fixed factor) yields $\underline{F}_2$. A significant $\underline{F}_1$ implies that the effect ought to be replicable with different subjects but the same materials; a significant $\underline{F}_2$ predicts replicability when the materials are changed but the same subjects are used. Clark presents a formula for combining $\underline{F}_1$ and $\underline{F}_2$ to produce min $\underline{F}'$ which predicts replicability when both new subjects and new materials are used.[27]

Correcting the "fixed-effect fallacy" however is not as simple as Clark implies. Past research cannot merely be reanalyzed. $\underline{F}_1$ and $\underline{F}_2$ can be combined to produce an interpretable min $\underline{F}'$ only if the power of each analysis is roughly similar. It is very likely that pre-Clark studies in which only an $\underline{F}_1$ data analysis was intended will have a sufficiently powerful $\underline{F}_1$ but a much less powerful $\underline{F}_2$. In such cases, the insignificant min $\underline{F}'$ is at least as likely to result from the lack of power in the materials design as from the inability of the effect to generalize.

The inclusion of a min $\underline{F}'$ analysis is something that should <u>precede</u> rather than follow the choice of a design. Clark is obviously aware of this where future research is concerned: "The most important rule to keep in mind . . . is this: An experimental design is only as sensitive as the less sensitive of the two subdesigns it contains" (1973, p. 349). Yet, in seeming self contradiction, he reanalyzes

---

[27] See Clark (1973) page 356 for an explication of the derivation of the formula for min $\underline{F}'$. See the Results and Discussion sections for an example of the use of min $\underline{F}'$.

some past research and concludes that the insignificant min $\underline{F}'$ is evidence that the effect will not generalize to other language materials.[28] By including case studies in which re-analyzed results are universally nonsignificant, Clark gives the unfortunate impression that both the quasi $\underline{F}$ and the min $\underline{F}'$ tests are extraordinarily conservative. He extravagently claims that "almost everyone" is committing the "fixed-effect fallacy" thereby implying that almost every study is a target for re-analysis and probably insignificant. Yet he fails to note that the insignificance which is indeed likely, is probably a function of the inappropriate nature of the re-analysis and so not really very informative about the nature of the effect. In this case, an insignificant min $\underline{F}'$ is like an insignificant $\underline{F}_1$ that describes a study that used only four subjects.

The min $\underline{F}'$ analysis is surely a more stringent test than just an $\underline{F}_1$, but there is no reason to believe that it is as intimidating as Clark's re-analyses would suggest. Forster and Dickenson (1976) used a Monte Carlo technique to compare Type I error rates for $\underline{F}_1$, $\underline{F}_2$, $\underline{F}'$ and min $\underline{F}'$ analyses when both subject and material variances were manipulated. Their results, which are summarized in multistage decision rules, suggest that under some conditions $\underline{F}'$ and min $\underline{F}'$ are

---

[28] In fact, Clark (1973, p. 349) points out that for one study that he re-analyzed, with 56 subjects and 8 materials, it is the much smaller $\underline{F}_2$ that places an upper bound on the $\underline{F}'$. But he does <u>not</u> conclude that therefore his re-analysis is inappropriate.

overly conservative but, for other conditions, $\underline{F}'$ and min $\underline{F}'$ are to be preferred over either $\underline{F}_1$ or $\underline{F}_2$ or both.

Empirical Solutions?

Wike and Church (1976) dispute the utility of $\underline{F}'$ and min $\underline{F}'$ arguing that they are only approximate $\underline{F}$ tests and that as such, statistical information about their properties is quite limited. Most investigations into the goodness of fit of the quasi-$\underline{F}$ approximations deal with conditions of independent observations. Thus, there is virtually no information about quasi-$\underline{F}$'s involving repeated measures which are very common in psycholinguistic research. Wike and Church conclude that, although a major aim of research is generality, Clark's statistical prescriptions for achieving generality are "unsound." The problems of induction cannot be solved by simply inserting random factors where fixed factors used to be. Rather, they "suggest that investigators continue using fixed factor designs about which more is known and seek nonstatistical generality by means of various modes of replication" (p. 254).

> . . . An investigator can replicate his findings with the same subjects and materials. He can replicate with new subjects or materials or both. Other investigators can replicate with different subjects and the same or different materials. Replication can also assume the form of what Lykken (1968) terms 'constructive replication' in which the reliability of an empirical fact is put to a test. Or by 'systematic replication' (Sidman, 1960) in which an investigator deliberately departs from the conditions of a previous experiment . . . in order to assess the generality of a relationship as well as its dependability (p. 254).

This empirical approach removes the burden of induction from the test of significance, and distributes the responsibility throughout the research community. The generality of an effect, then, is inferred by examing its stability and its lack of stability over the variable conditions of replication. This position seems to be a concrete realization of Popper's (1962) conception of the manner in which scientific progress is attained: bold conjectures are offered to a scientific community for refutation. But, within the realm of psychological experimentation and statistical analyses, what is the means of refutation? The analysis of variance model gives the researcher two ways of describing results: the null hypothesis was rejected, and one may draw some conclusions about the difference between the treatment groups; or 2) the null hypothesis was not rejected, and no conclusions are warranted regarding the equality or inequality of the treatment groups. The possibility of accepting the null hypothesis is, strickly speaking, not an option within the model because of the inability to distinguish between the actual equality of groups and the apparent equality of groups due to measurement error. By stepping outside the model and comparing the power of various experiments, a researcher may indeed be relatively sure that the failure to reject the null hypothesis represents a lack of an effect rather than a lack of power. But then there is the added problem of publicizing that information. Greenwald (1975) cogently describes the editorial prejudice encountered when trying to publish

null results.

A purely empirical solution to the problem of induction is clearly inadequate. In addition to the statistical problems involved in accepting null results, and the pragmatic problems involved in publishing null results, there is also the theoretical problem of acknowledging and interpreting null results. A well-know "attempt-to-replicate" will be briefly examined in order to demonstrate the futility of replication if that replication is not embedded within an explicit theoretical framework.

Rosenthal (e.g., 1966) has held that the Experimenter Bias Effect--the tendency of experimenters to unwittingly twist their results in the direction of their predictions--is pervasive and varied. He has, he claims, produced the phenomenon in dozens of published studies, with arenas as diverse as rat laboratories and classrooms. Rosenthal has held, moreover, that a great many experiments, including a number of very seminal studies, may be invalid.

A number of psychologists take issue with Rosenthal's findings and conclusions. Chief among these is Barber who has sought to demonstrate that the Experimenter Bias Effect cannot be replicated and hence does not exist (Barber, et al., 1969). In "five attempts to replicate the Experimenter Bias Effect," in which nearly all the suggestions of Wike and Church were employed, Barber and his associates could replicate Rosenthal's results in none of them. Rosenthal, in a rejoinder, denied that his experimental conditions had been

replicated by Barber; and he also claimed that, if Barber had used other statistical tests, he too would have "observed" the Experimenter Bias Effect in his own laboratory. Barber, in turn, commented on the problem associated with "post-mortem analyses." In short, Barber followed the procedure recommended by Wike and Church not one but five times (a point emphasized by Barber), and yet Rosenthal's conclusions regarding the Experimenter Bias Effect remained unchanged.

The exchange between Rosenthal and Barber might continue indefinitely, yet the outcome is quite predictable: Rosenthal will remain convinced that the Experimenter Bias Effect is a real phenomena, and Barber will remain sure that he has demonstrated its nonexistence. But how can that be? How can two diametrically opposed views be "supported" by the same "evidence"? In a commentary on the Rosenthal and Barber exchange, Levy (1969) suggests that this collective monologue is a predictable outcome of experimentation that is more procedural than theoretical. Levy broadens the scope of the controversy and questions the value of the replication process itself:

> The perfect replication is a fiction, and I shall take the heretical position that this is just as well . . . For obvious reasons, no experiment can ever duplicate another in every detail, and so this question (of replication) turns on whether the variations between (proposed replications are) trivial or important. . . . this choice . . . requires either a theory which states the parameters involved in (an experimental effect) or a body of systematic research from which these parameters might be induced . . . Thus whether the findings of Barber et al. can be taken as contradictory to those of Rostenthal . . . is a moot question, and this, I would suggest will be found true wherever replications are

> attempted of experiments dealing with phenomena which are not embedded either within some theoretical framework or extensive body of systematic research (p. 15).

If, however, the phenomena to be investigated are part of an explicit conceptual formulation, then salient characteristics of the sampled populations can be delineated. Results from experiments in which samples are drawn explicitly from the same populations are of interest and experimentation with other populations--whose characteristics are explicitly stated--can then provide some information about the generality of the results.

Hume's Paradox--No Solution

Generality is one of the major aims and problems of experimental psychology. To generalize is to make an inductive leap. Statistical models and empirical prescriptions for replication have attempted to automate and legislate that leap. In this regard, they have failed, though no doubt statistical and empirical evidence is quite useful to the researcher compiling a case for an inductive inference.

Hume has classically stated the paradox of induction --it is illogical and it is unavoidable:

> The idea of a necessary connexion among events arises from a number of similar instances which occur in the constant conjunction of these events . . . There is nothing in a number of instances, different from every singly instance . . . except only, that after a repetition of similar instances, the mind is carried by habit, upon the appearance of one event, to expect its usual attendant, and to believe that it will exist. This connexion, therefore, which we <u>feel</u> in the mind . . . is the sentiment or impression <u>from</u> which we form the idea of power or necessary connexion. Nothing farther is in the case (II, vii).

In experimental psychology, those "feelings" of certainty arise from compelling arguments that make use of statistical, empirical and rational evidence. Ultimately, such arguments derive their force from the theoretical sense that they make rather then simply from the extent to which they adhere to certain formal content-free inductive techniques.

Focus of the Present Study

The study to be presented examines the effect of clausal analyses on the storage and retrieval of information presented in a specific class of sentences. In addition, the lexical and surface structure cues that relate to the clause structure of the sentence are considered. The experimental paradigm used to study these issues is the probe paradigm.

Experimental Psycholinguistics: The Probe Paradigm as a Tool for Investigating Linguistic Structures and Perceptual Strategies

In a probe-latency experiment, a sentence is presented to a subject. The sentence is immediately (i.e., within 50 msec) followed by a word that may or may not have been in the sentence. The subject's task is to indicate as quickly as possible whether or not that word (the probe) appeared in the sentence. The dependent measure is the reaction time.

Presumably, in order for subjects to perform the rather simple task they must search a stored representation of the sentence after the probe word is encountered. If

clausal analysis is an important procedure in sentence interpretation, then it is a reasonable corollary to suppose that a sentence is stored clause-by-clause. But, until the decision of what constitutes a clause is made, the elements of the sentence will be held in an immediate memory buffer. The buffer is emptied when it contains a complete clause, and the clause as a whole is stored in short-term memory. Elements from a common clause are stored together and separated from elements of another clause. If a search of short-term and immediate memory is undertaken, the prediction is that the search will also proceed clause-by-clause, and that information stored in earlier clauses is less accessible than information from more recently stored clauses.[29] The difference in accessibility of material in earlier and later clauses is experimentally examined by comparing reaction times to probes from each clause.

Caplan (1972) constructed pairs of sentences which shared common lexical material. As in the Garrett et al. (1966) study, the common material was cross-recorded to insure that subjects would hear the same local acoustic cues with each version. For all sentence pairs, the first common word was, in one member, the last word before the clause boundary, while in the other pair member, the word came directly after the clause boundary. For example:

---

[29]Note that as in the click studies, this research is concerned with the psychological reality of clause structures, not the processes that underlie the use of such structures.

22a. When the sun warms the earth after the rain, clouds soon disappear.

22b. When a high pressure front approaches, rain clouds soon disappear.

If linguistic material is stored and retrieved in a clause-by-clause fashion, then the accessibility of the word "rain" ought to increase from 22a to 22b since "rain" will be stored as part of the first clause in 22a and part of the second clause in 22b. Results supported the prediction; probe words from the final clause consistently provoked shorter reaction times than probes from the first clause.

It is tempting to infer from this study that sentences are stored clause-by-clause and that this implies a last-in-easiest-out retrieval process. But the results equally support the following description: sentences are stored clause-by-clause, and a labeling of the structural relationship between the clauses is also stored; information from the main clause is more accessible than information from the subordinate clause. Here, in addition to a clausal analysis, a surface-structure analysis utilizing heuristic B would contribute to the way in which grammatical material is stored. And the retrieval process would be sensitive to the storage method. Heuristic B is easily applicable to all the sentences in the study since each initial subordinate clause is introduced by an adverb that clearly marks the embedding. Therefore, since all of the sentences that Caplan used were of the form: subordinate clause/main clause, there is no way of

deciding between the alternative descriptions.

By utilizing both subordinate/main and main/subordinate type sentences, as well as coordinate sentences that could be described as main/main, Kornfeld (1974) attempted to assess the independent contributions of the temporal order of the clauses and the dominance relationship between the clauses. The effect of temporal order was strongest for coordinate sentences which by definition show no dominance relationship between clauses. For complex sentences with adverbial clauses, both clause recency and dominance effects were present, but the nature of their combined effect was relatively unstable. In one experiment, the variables operated in a more or less additive fashion. That is, probes from first clauses that were subordinate evoked the longest reaction times while probes from second clauses that were main evoked the shortest reaction times. But, in another experiment, the dominance variable overrode any reaction time decrements due to recency.

Note that the conflicting experiments used different types of lexical items in the stimulus sentences. More specifically, in the experiment where dominance was the overriding variable, relative and complement clause types were investigated as well as adverbial clauses. Thus, in an effort to keep the multiple versions of each sentence as alike as possible, many of the sentences with adverbial clauses also contained markers that could dominate subject or object complement clauses. These complement markers were

either main verbs, head nouns, modal auxilaries, or adjectives. In the other experiment, where dominance and recency were additive, only sentences containing adverbial clauses were used. Thus, the incidence of complement markers was extremely low. Recall that in some situations (Garrett et al., 1966) the perceptual complexity of a sentence is increased by the inclusion of a verb or other lexical marker that is compatible with multiple constructions. Thus, it is not too surprising that the results from an experiment with complex lexical items do not mirror the results from a similar experiment with simpler lexical items. This recency vs. dominance or recency plus dominance issue is still unresolved.

In addition, it is obvious that there are other variables besides clause structure that contribute to the reaction time, such as the length of the sentence, the distance from the probe word to the end of the sentence, the type of response (vocal or manual) that a subject must give, etc. The prediction is that reaction time differences will be reflections of clause boundary differences--<u>other things being equal</u>. Lang (1974) found that it is not possible to compare sentences that differ by four or more syllables either in length or in probe distance. Such differences in the external characteristics of the sentence can mask the effects of structural differences.

Experimental Psycholinguistics: Clausal, Lexical, and Surface-Structure Analyses

In order to use the probe paradigm to investigate structural aspects of sentence processing, attention should be paid not only to the major structural variable of interest (in this case clause boundaries), but also to the lexical and surface-structure features of the sentence that may ease or complicate the interpretation of the sentence, and to the external characteristics of sentences to be compared. In the study to be presented, this is accomplished in the following manner:

Manipulated variables:

A. Temporal order of the clause (a perceptual variable)--the probe is either from the first or second clause of the sentence.

B. Type of clause (a structural variable)--the probe is either from the subordinate or main clause of the sentence.

Controlled variables:

A. Although only complex sentences with adverbial clauses are considered, the lexical items of each sentence readily allow for the construction of complement clause versions of each sentence.

B. All initial subordinate clauses contain surface-structure cues that mark the embedding.

C. Only sentences with identical external characteristics (i.e., length of the sentence and probe distance in

syllables) are compared.

Moreover, both subjects and language materials are conceptualized as random factors. That is, the experimental effects are of interest because it is assumed that they are not restricted to the particular subjects and materials sampled in this experiment. The data analyses will therefore include the computation of Clark's min $\underline{F}'$ statistic. This statistic is computed by combining information from a subjects analysis ($\underline{F}_1$) and a materials analysis ($\underline{F}_2$). As a preliminary step toward equating the power of the $\underline{F}_1$ and $\underline{F}_2$ analyses, the number of subjects and sentences used in the experimental design will be equal.

Experimental Psycholinguistics: Methodology

In psycholinguistic research, a particular linguistic phenomenon is operationalized as a set of sentences that conform to some explicit structural description. Those sentences form a rather homogeneous set of experimental stimuli. Researchers formulate hypotheses regarding the specific influence of the linguistic variables, but the testing of those hypotheses involves more than just distributing the stimuli to subjects, observing responses, and making causal inferences. Psycholinguists, like all other experimental psychologists, must also grapple with difficult problems of experimental invalidity.

Campbell and Stanley (1966) have provided a concise and influential framework within which to assess the validity

of an experiment; they have recommended that the concept of validity be divided into two parts--internal and external. An experiment is internally invalid if experimental and control groups differ systematically along a dimension other than the independent variable. Such confounding makes it difficult for the researcher to infer that a causal relationship exists between the independent and dependent variables in the experiment. Internal validity is the first requisite of any study. If an experiment is internally valid--if, that is, the experimenter exerted appropriate and powerful control over the variables relevant to his/her results--then he/she will wish to make some broad statements about the generality of the phenomenon just observed in the laboratory. An experiment is externally invalid to the extent that the sample or laboratory situation is an inaccurate copy of the population or "real world" to which one wishes to generalize the results.[30]

When sets of linguistic stimuli are constructed for an experiment, the aim is to make the sets alike on all relevant variables except one--the independent variable. Determining which variables are and are not relevant is in part an intuitive decision that researchers make based on theoretical information and research experience. Perhaps the most

---

[30]One way to conceptualize the difference between factors that jeopardize the internal and external validity of a study is to consider how the effects of those factors would be discussed within the analysis of variance model. Factors affecting internal validity add alternative main effect statements to the design; factors affecting external validity necessitate the addition of interaction statements to the design.

important function of the scientific community is to critically evaluate a study--to determine whether there are plausible alternative explanations of the results. An experiment that is relatively internally valid, then, is one in which potentially confounding variables have been isolated and successfully controlled.

When a number of experiments are conducted in sequence, using the same experimental paradigm, the result is normally a substantial increase in internal validity. For example, Fodor and Bever (1965) have suggested that errors in subjectively locating a click in a sentence are predictable on syntactic grounds: that is, constituents such as phrases or clauses show a resistance to interruption and so clicks tend to be displaced to the constituent boundaries. But, for the stimuli used, there is an alternative explanation for the displacing. There is a longer acoustic pause at constituent boundaries than at other places in the sentence. Perhaps it is the pause, per se, that is attracting the click, rather than any more abstract properties of the sentence. In fact, the authors report that Garrett, in an unpublished experiment, has demonstrated such an effect with random digits. In order to eliminate the confounding of the syntactic and acoustic variables, Garrett, et al. (1966) constructed pairs of sentences like the following:

9a. (In her hope of marrying) Ann was surely impractical.

9b. (Your hope of marrying Ann) was surely impractical.

Since the common lexical items are recorded only once and then spliced onto the initial parts of each sentence, the prosodic cues of each sentence are identical. But the syntactic structure of each sentence differs. The independent variable of interest is not confounded with the acoustic variables. The Garrett et al. study is more internally valid than the Fodor and Bever (1965) study. Similarly, Caplan (1972) has performed a series of probe latency studies and reports that probes from the first clause evoke longer reaction times than probes from the second clause. Early studies compared sentences like the following:

23a. No matter how carefully you scheme, crime won't pay.

23b. Whenever this goalie stops the puck, fans go wild.

The independent variable--probe position--is, in this case, confounded with differences in lexical material stress, intonation, rhythm, etc. In order to control for these confounding variables, Caplan followed a procedure consistent with Garrett et al.'s reasoning. Sentence pairs were constructed that shared common lexical items but had different clause boundaries, as in 22a and 22b:

22a. When the sun warms the earth after the rain, clouds soon disappear.

22b. When a high pressure front approaches, rain clouds soon disappear.

Controlling for lexical difference increased the internal validity of the experiment. Kornfeld (1974), however, pointed out that additional confounding still remained. Although first-clause probes continued to evoke longer reaction times than second-clause probes, the effect might be due to the clause-boundary effect, or to the type of clause that included the probe (the first clause was always subordinate, the second clause was always main), or to the syntactic category of the probe word (probes in the first clause were nouns, in the second clause adjectives). By incorporating both subordinate-main and main-subordinate orders, and by limiting the syntactic category of all probe words to nouns, Kornfeld further reduced the confounding and thus increased the internal validity of the study.

In summary, the internal validity of an experiment increases as the researcher more accurately defines the stimuli used and thereby decreases the possibility that an alternative description will account for the experimental results. For the present study, the simulus sentences have been constructed and defined in such a way as to eliminate a number of rival explanations of the experimental effects:

1. The definitional advances of Kornfeld (1974) have been retained: subordinate-main and main-subordinate clause orders are used; and all probes are nouns.
2. Although only complex sentences with adverbial clauses

are considered, the lexical items of each sentence readily allow for the construction of complement clause versions of each sentence.

3. Since differences in the external characteristics of sentences (number of syllables in the sentence, probe distance length in syllables) can produce differences in reaction times, all sentences were identical with respect to these external characteristics.

Thus there is good reason to expect that any results obtained will be due to the manipulation of the variables of interest (probe position and type of clause) although additional refinement of the paradigm must inevitably follow.

The Inclusion of Filler Sentences. Once a well-defined set of sentences has been constructed to relatively unambiguously embody a particular independent variable, then the external validity of the study may be considered. In assessing the degree to which the experimental or laboratory findings might be generalized to real situations two related concerns arise:

1. The problem of artificiality: are sentences constructed for a laboratory experiment like sentences spontaneously used in the real world?
2. The problem of reactivity: does the subject's awareness of being measured alter the processes involved in sentence interpretation?

Both problems are, in an absolute sense, insoluble. Most experimental sessions are quite unlike any non-laboratory

situation: tones and buzzers precede and follow sentences, the sentences are not related to any topic of discourse, and some sort of question is asked after every sentence. Subjects are quite aware that they are participating in an experiment and are hooked up to some type of reaction-time device.

Still there are aspects of experimental design that can be used to attenuate though not eliminate the artificiality and reactivity. In the present context, filler sentences (sentences that do not meet the requirements defined for the experimental sentences) should be included in any study to increase its external validity. Filler sentences are structurally more heterogeneous than the experimental sentences. (In the present study, all experimental sentences contain two clauses, one of which is adverbial; the fillers are not so restricted, and include coordinate clauses which themselves may contain infinitivals and other subordinate structures.) With the inclusion of filler sentences, the array of sentences presented to subjects is not the repetitive structural array that would result from just presenting the experimental sentences.

More importantly, the heterogeneity of filler sentences disguises the intent of the study. Orne (1962) has suggested that when reactive measures are used, researchers must consider the active problem-solving characteristics of their subjects, and their propensity to be "good" subjects. The reactions of subjects are always joint functions of both

the experimental manipulation and the hypotheses subjects formulate about the nature of the investigation. If experimental sentences are used exclusively, then it may become relatively easy for subjects to figure out the expected response and produce it. For example, with the probe paradigm, after hearing a number of the experimental sentences, subjects might easily surmise that the best strategy for achieving short reaction times would be to signal "IN" as quickly as possible after the probe word. The experimental task would, in that case, have changed from sentence comprehension to monitoring, and the new task is unrelated to the purpose of the experiment--to investigate the structures involved in sentence processing. But, if the experimental sentences are interspersed among a set of filler sentences that are structurally different from each other and from the experimental sentences, and if those fillers are followed by probe words that are sometimes not from the sentence, then it is quite unlikely that subjects will form uniform and correct predictions about what they should do on the succeeding sentences. In addition, subjects were told initially that periodically they would be asked to paraphrase a sentence just heard. Thus, a number of precautions helped to insure that subjects listened to each sentence for its meaning.

Despite their name then, filler sentences serve a critical methodological function in psycholinguistic research. Their inclusion increases the external validity of the study,

thereby making more legitimate the generalizations that experimental researchers make about linguistic abilities.

The Analysis of Filler Sentences. Though the use of filler sentences in a psycholinguistic experiment is an all but automatic procedure, the analysis of data from those sentences is non-existent. Data from filler sentences is characteristically collected from subjects for whom each sentence is as important as any other sentence. Data from filler sentences is ignored by researchers for whom such data is useless and uninteresting, a half-hearted and unexamined attempt to control some "nuisance" variability.

Lang (1974) has suggested, however, that data from filler sentences provides the researcher with the unique opportunity to verify empirically the implicit and common assumption that the random assignment of subjects to experimental groups results in groups that are initially equal and interchangable. The filler sentences can be conceptualized as a variable of "no treatment." If it can be shown that when groups of subjects are treated alike there are no resulting differences between the groups, then differences between experimental groups can be more strongly attributed to the treatment variables. A filler analysis that supports the hypothesis of the initial equality of the groups adds a measure of internal validity to the study. The filler analysis, that is, gives the researcher added confidence that the samples are indeed equivalent due to random sampling, and that the experimental instructions were interpreted correctly,

without systematic bias.

There is an alternative conception of the role of a filler analysis. Filler sentences may be conceived of as a variable that measures "context" effects. Although each group will receive the same filler sentences and probes, the other sentences that each group hears will not be identical. Each group will hear sentences that uniquely represent specific levels of the independent variables to be studied. The differences in those sentences could lead to the formation of different response sets across the different subject groups. The filler analysis provides a check on this.

In the present study, filler sentences are an integral part of both the experimental procedures and the data analyses.

Experimental Psycholinguistics: Exploration

It has been shown that probe tasks are sensitive to the grammatical structure of linguistic material (Caplan, 1972; Fodor et al., 1974). What of the sensitivity of the probe task to non-grammatical variables? Specifically, the following questions are asked:

1. When adjacent words from the same clause are independently probed, will the temporal order of the words be reflected in reliable reaction-time differences?

If the temporal order of linguistic material <u>within</u> a clause follows the same pattern of storage and retrieval as do multiple clauses then the answer will be "yes." If the

syntactic relationships among words within a clause interacts with the recording of the temporal order of the words, then the answer will be "no," or at least "not necessarily."

2. When the probe word is not from the sentence, will probes that are phonetically related to a word in the sentence evoke reliably different reaction times than probes that are neither from the sentence nor related in any way to words in the sentence?

Preliminary research by Limber and Lang (1975) suggested that probes which rhyme with a word in the sentence are particularly troublesome to individuals learning English as a second language. It may be that this confusion effect of rhyming probes is limited to individuals learning the language, or it may be a general characteristic of all language users, with second-language learners and native speakers differing only in the magnitude of this confusion effect.

In a preliminary attempt to address these issues, half of the filler sentences from the present study are used with a number of different probe types. Specifically, adjacent words from the same clause are independently probed; and probes that rhyme with a word in the sentence are compared with probes that do not.

Goals of the Present Study

1. To define the relationship between the perceptual variable of temporal order of clauses and the structural variable of type of clause for a restricted class of

of sentences (complex sentences with adverbial clauses containing syntactically complex verbs).

2. To examine the implications that min F' analyses have on the design of psycholinguistic experiments.
3. To illustrate the value of filler sentences in the design and analyses of psycholinguistic research.
4. To explore the potential of the probe paradigm as a tool for investigating the importance of the temporal order of words within a clause.
5. To investigate the effects of various probe types on native English speakers.

METHOD

Variables and Hypotheses

Design 1

The probe paradigm was used to investigate the clausal analyses that underlie the storage and retrieval of linguistic information in a class of complex sentences. Specifically, the independent and interactive effects of two variables were studied: the temporal order of the clauses and the dominance relationship between the clauses. Temporal order was investigated by probing a word from either the first or the second clause; dominance was indexed by the type of clause that the probe was in--subordinate or main.

The design may be summarized as follows:

		Type of Clause	
		Subordinate	Main
Probe Position	First Clause		
	Second Clause		

Critical sentences were adapted from those used by Kornfeld (1974). There were four versions of each sentence. Each version conformed to the specific description of one cell in the design. In the following diagram, each line represents a clause and is labeled either subordinate (S) or

main (M), and "x" represents the position of the probe:

version a.	S x,	M
b.	M ,	x S
c.	M x,	S
d.	S ,	x M

Although only complex sentences with adverbial clauses are considered, the lexical items of each sentence readily allow for the construction of complement clause versions of each sentence. Thus, these sentences contain a high incidence of markers that can dominate subject or object complement clauses. In this way, the present findings would be useful for comparative purposes for future research dealing with non-adverbial constructions. Lexical and semantic differences within each sentence set were minimized. The same probe word was used in all versions of a sentence set, and the syntactic category of that probe was constant across all versions. The sentence length and probe distance in syllables were identical for all versions of all sentence sets: all sentences were 24 syllables long, and the probe was always 12 syllables from the end of the sentence.

Twenty sentence sets were used. The following set illustrates how the above restrictions combine to form a sentence set:

Critical sentence #1, probe word LAND

version a. When the greedy rancher began purchasing land, farming plots became quite scarce in Suffolk county.

b. The greedy rancher tried to purchase farms when land was becoming more and more scarce in Suffolk county.

c. The greedy old rancher tried to purchase more land when farms were becoming scarce in Suffolk county.

d. When the greedy old rancher began purchasing farms, land quickly became quite scarce in Suffolk county.

A complete list of sentence sets appears in Appendix A.

If, all things being equal, material from second clauses was more accessible than material from first clauses, then there should be an overall main effect for the probe position factor; and the mean for probe position - 1 ought to be significantly greater than the mean for probe position - 2.

If, all things being equal, material from main clauses was more accessible than material from subordinate clauses, then there should be an overall main effect for the type of clause factor; and the mean for subordinate clauses ought to be significantly greater than the mean for main clauses.

If the independent effects of both variables were additive (i.e., there is no significant interaction), then the four cells in the design ought to be rank ordered in the

following way: the cell with version (a) sentences ought to have the largest mean reaction time; the cell with version (d) sentences ought to have the shortest. The (b) and (c) version cells ought to be intermediate--the particular rank ordering depending on the relative strengths of the two independent variables.

Figure 3 represents the hypothesized relationships among the cell means ("+" indicates a relative increase in reaction-time latencies; "-" indicates a relative decrease).

		Type of Clause	
		Subordinate	Main
Probe Position	First Clause	+ First Clause + Subordinate	+ First Clause - Main
	Second Clause	- Second Clause + Subordinate	- Second Clause - Main

Figure 3. Hypothesized rank ordering of the cell means from design 1.

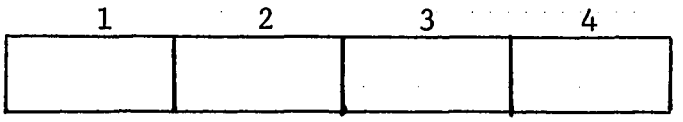
Design 2

In order to decrease the liklihood of subjects forming a response set to the homogeneous set of critical sentences, twenty filler sentences were constructed and integrated with the critical sentences. All filler sentences were complex sentences, 24 syllables long. To discourage position sets the probes from 14 fillers were 6 or 18 syllables from the end

of the sentence. To discourage subjects from automatically responding "in" to each probe, six other fillers had probe words that were not found in the sentence. A complete list of filler sentences appears in Appendix B. The sentences on the list are grouped according to type of sentence and temporal position of the probe.

Filler sentences were thus used to mask the intent of the study. In addition, the data from filler sentences can be used to test the implicit assumption that differences between groups (in design 1) were due to treatment differences rather than to either any intrinsic differences between the groups per se, or to differences that might be due to response sets that different combinations of experimental sentences might encourage. In the present study, the performance of the four groups of subjects from the previous design was examined under the condition of no treatment. That is, all four groups received the same treatment--the same 20 filler sentences and the same probes. The design can be outlined as follows.

Subject Groups



Group 1 contained the subjects who received version (a) of the critical sentences; group 2, version (b); etc.

Design 1 thus examined subject groups that were treated differently, while design 2 looked at those same subject groups when they were treated alike. The hypothesis was that there would be no significant differences between the groups in design 2. Randomly selected groups treated alike would respond, on the average, alike. Further, the contextual differences between the groups would not be significant.

Design 3

Twenty filler sentences with the same external characteristics as those used in design 2 were constructed. For each of these sentences, however, four different probe words were used: IN, OUT, RHYME, and ADJACENT-IN. The probe in the IN group was a word from the sentence that was either 18, 12, or 6 syllables from the end of the sentence. In the OUT condition, the probe was not from the sentence and had no relation to any word from the sentence. A word that rhymes with the IN-word was the probe in the RHYME condition; and the word immediately preceding the IN-word was probed in the ADJACENT-IN condition. A prototypic example:

> Because the tall young man was so strikingly handsome, he looked terrific in all kinds of clothing.
>
> probes: (IN) (OUT) (RHYME) (ADJ.-IN)
>
> MAN JOB PAN YOUNG

A complete list of the filler sentences and the four probes appears in Appendix C. Again, the sentences are grouped by sentence type and probe position.

If the temporal order of linguistic material within a clause followed the same pattern of storage and retrieval as do multiple clauses, then the mean reaction time for the ADJACENT-IN group ought to be greater than the mean reaction time for the IN group. If deciding that a probe was not from a sentence required a search of both clauses in the sentence, then the mean reaction time for the OUT group ought to be longer than either the IN or the ADJACENT-IN group means.

Preliminary research by Limber and Lang (1975) suggested that, for individuals learning English as a second language, probes which rhyme with a word in the sentence evoked longer reaction times than either probes from the sentence or probes that are not from the sentence and do not rhyme with any word in the sentence. The RHYME group was included in the present study to investigate the magnitude of that confusion effect in native speakers. It was hypothesized that the mean reaction time for the RHYME group would be the largest of the four.

Probes in the IN group were either 18, 12, or 6 syllables from the end of the sentence. ADJACENT-IN probes were also distributed throughout various positions in the sentence. Thus, if reaction time to a probe was related to the temporal position of that word in the sentence, then the reaction times that contribute to both the IN and ADJACENT-IN means ought to be relatively diverse. Conversely, for the OUT group, the task was very nearly the same for each sentence: to make a complete search of a 24-syllable sentence. The

variance of the OUT group should be the smallest of the four; the variances of the IN and ADJACENT-IN groups should be larger, and roughly comparable to each other. The variance of the RHYME group should be intermediate: smaller than either the IN or ADJACENT-IN group variances because the RHYME task too was very similar each time; but larger than the OUT group because the probes were phonetically related to words that were distributed throughout the sentence.

Each subject group as defined in designs 1 and 2 received one type of probe. Thus, the design was as follows:

Type of Probe

IN	OUT	RHYME	ADJACENT-IN
Group 1 Sx, M	Group 2 M, xS	Group 3 Mx, S	Group 4 S, xM

In a strict sense, designs 1 and 3 were confounded: each subject group was exposed to a unique combination of variables from both designs. Yet, other aspects of the design mitigated the confounding. It might be argued that the addition of design 3 encouraged the four subject groups to develop different response sets to the sentences: for groups 1 and 4, only 6 out of 60 times during the experiment was OUT the correct response; while for groups 2 and 3, OUT was a correct response 26 out of 60 times. Those IN/OUT percentage differences could have been responsible for design 1 differences. But, any response sets that affected design 1 reaction times

should also have affected design 2 reaction times. Thus, the analysis of data from design 2 becomes an important measure of the "context" effect. Further, since the IN/OUT percentages for groups 1 and 4 were identical, their "response sets" should have been quite similar. Yet, recall that for design 1, those two groups were predicted to be the largest and the smallest of the four.

Tapes and Equipment

Four tapes were used in the present experiment. Each tape contained 60 sentences (20 critical sentences, 20 fillers from design 2, and 20 fillers from design 3). Only one version of each critical sentence and only one probe type for the filler sentences from design 3 was used in each tape. The fillers and probes from design 2 were identical on all tapes. The make-up of the four tapes is summarized in Figure 4. The order of the 60 sentences was randomly determined and was the same for all four tapes. A chart of the random sentence order appears in Appendix D.

The sentences were recorded on a Dokorder 7140 4-track recorder. In order to minimize intonation and stress effects, the sentences were read by a male speaker in a neutral and steady manner. After each sentence, a 500 Hz 50 msec low-intensity tone was spliced onto the tape to indicate the end of the sentence to the subject. This was necessary to prevent the probe from being interpreted as part of the sentence (Caplan, 1970). After the tone, 100 msec of leader tape was

Tape A	Tape C
20 criticals--version (a) S x, M	20 criticals--version (c) M x, S
20 fillers from design 2	20 fillers from design 2
20 fillers from design 3-- version (a) IN probes	20 fillers from design 3-- version (a) RHYME probes
Tape B	**Tape D**
20 criticals--version (b) M , x S	20 criticals--version (d) S , x M
20 fillers from design 2	20 fillers from design 2
20 fillers from design 3-- version (b) OUT probes	20 fillers from design 3-- version (d) ADJACENT-IN probes

Figure 4. Schematic outline of the sentences that appeared in each of the four tapes used in the experiment (designs 1, 2, and 3).

added. The probe word directly followed the leader tape. A high-frequency noise burst (5000 Hz) located on a second channel, and inaudible to the subject, activated a msec timer at the onset of the probe word. The subject's response stopped the timer. The delay between sentences was two seconds. Another 500-Hz 50 msec low-intensity tone preceded the next sentence. This pattern, illustrated in Figure 5, was repeated for each sentence.

Although only one channel of the tape was audible to the subject, the earphone plug was wired so that the subject heard that one channel in both ears.

Movement of a balance toggle-switch to one of two positions stopped the msec timer.

Subjects

Sixty-eight undergraduates from Grand Valley State Colleges were used. All subjects were enrolled in an introductory psychology class and participated in the experiment for extra credit. All were right-handed native English speakers without any history of hearing difficulties.

Subjects were randomly assigned to one of four groups. There were 17 subjects in each group. Subjects in group 1 listened to tape A; group 2, B; etc.

Procedure

The subjects were told that they were participating in an experiment in auditory perception. They were to listen to the sentences presented to them and to indicate whether

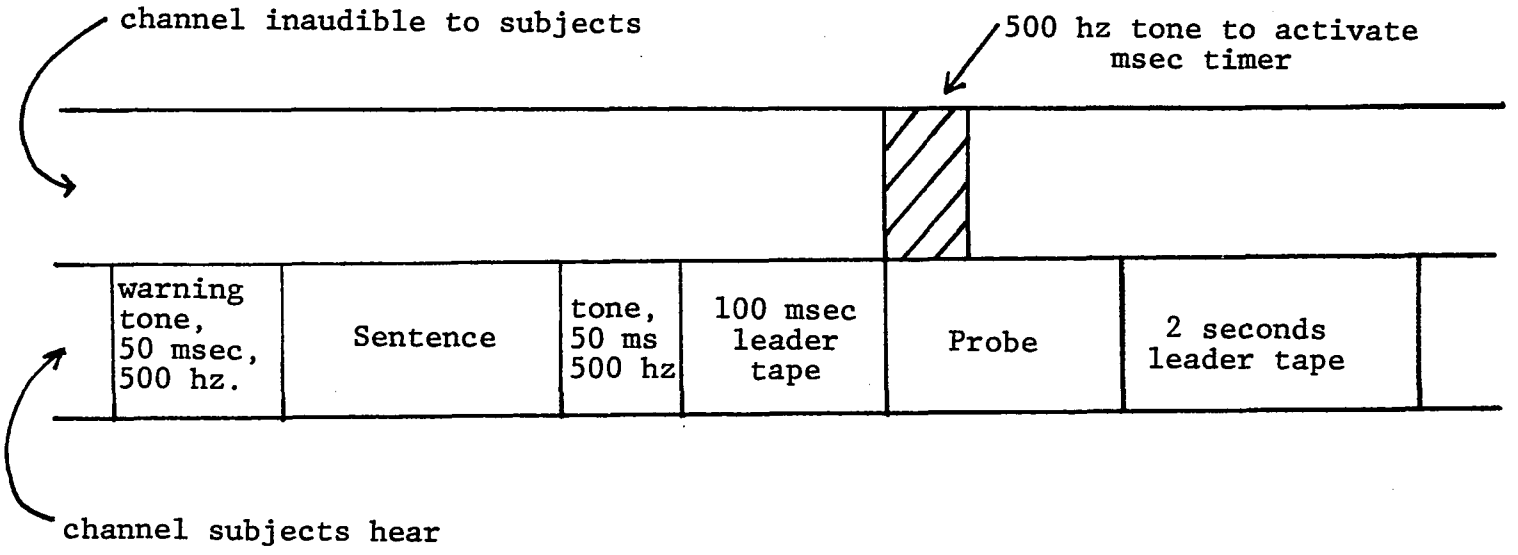


Figure 5. Schematic outline of the arrangement of sentences, probes, and tones on the experimental tapes.

whether they thought the word coming immediately after each sentence had been in the preceding sentence. If they thought the probe had been in the sentence, they were to push the toggle switch to the position marked IN; if the probe was not from the sentence, the toggle switch was to be pushed to the OUT position. All subjects responded on all trials with their right hand. Subjects were instructed to make their responses as quickly as possible. They were permitted to correct any errors they thought they had made before the presentation of the next sentence. Incorrect responses that were immediately corrected were counted; all other incorrect responses were discounted. In addition, three times during the experiment, subjects were asked to paraphrase the previous sentence. The paraphrase trials were randomly determinded for each subject. The text of the instructions is presented in Appendix E. Reaction times and errors were recorded by the experimenter.

Reduction of the Number of Stimulus Sentences Analyzed

Although 20 sentences were constructed for each design, data from only 17 sentence was utilized in each analysis.

In design 1 (the critical sentences) for sentences 4, 9, and 12, the time did not reliably start at the beginning of the probe word. Thus, the reaction times of subjects' responses could not be accurately recorded.

In design 2 (identical fillers and probes for all subjects), data from sentence 7 was not included in the analysis.

7. Their new furniture was delivered and they were surprised by how beautiful the room now looked.

WAS

Of all the sentences used, criticals and fillers, this sentence was the only one in which subjects erroneously said the probe was not in the sentence and did not then spontaneously correct their mistakes. Moreover, for those subjects who responded correctly, the reaction times were inordinately long. The mean reaction time for sentence 7 across all groups was 1214 msec as compared with a 662 msec mean reaction time across all groups for the 17 fillers. The problem seems to be that the second clause contains the word WERE, which because of its relation to WAS complicates the task. Although subjects from all four groups performed similarly on sentence 7, that sentence was not included in the analysis because, when the erroneous responses were dropped, that sentence contained 12 missing cells. In order to equate the number of sentences analyzed in design 2 with the number in design 1, two additional sentences were dropped. Sentences 2 and 6 were selected from a random draw.

In design 3 (same sentence, different probe word for each subject group), sentences 23 and 31 were dropped because two of the four probe words were inadvertently recorded on the wrong tape. Data from sentence 25 was also eliminated from the analysis.

Martha likes to walk three miles every day, but sometimes when it is snowing or raining she does not walk.

THREE WALK FREE SOX

Only after a majority of the subjects were run was it noticed the probe word for the ADJACENT-IN condition (WALK) appears in the sentence twice.

Missing Data

Each design was made up of a data matrix of 1156 cells: four groups of 17 subjects responded to 17 sentences. Because of occasional equipment failure or incorrect subject responses (those incorrect responses that were not simultaneously corrected), a small number of cells in each design (11 in design 1; 14 in design 2; 12 in design 3) have been estimated by using the following formula:

$$\bar{g} + (\bar{P}_i - \bar{g}) + (\bar{S}_j - \bar{g})$$

In design 1, if subject #i did not respond to sentence #j, then that data point would be estimated as follows: $\bar{g}$ is the mean of all subjects on all sentences for the group in which the missing cell appears; $\bar{P}_i$ is the mean of all the sentences #i did not respond to; and $\bar{S}$ is the mean of all other subjects on sentence #j. The missing point is thus estimated by the group mean adjusted for the particular contributions that might be expected from the subject and the sentence independently. The procedure does not take into account any unique subject x sentence contribution. By assuming that

contribution to be zero, this estimation procedures tends to reduce the error variability. However, since the number of missing cells estimated in each design is never more than 2.17 percent of the total number of cells in the design, the effect of the estimation procedure is presumed negligible.

RESULTS

Analysis of Critical Sentences (Design 1)

When the variables of Probe Position (first clause, second clause) and Type of Clause (subordinate, main) were crossed, four experimental groups resulted. The data from each experimental group was combined into one mean reaction time score. The mean of each experimental group as well as marginal means and the grand mean are presented in Table 1. These means were obtained by averaging across sentences and then subjects or equivalently, across subjects and then sentences. Either order of averaging necessarily produces indentical mean reaction times, but the variability associated with the mean changes with the variability of the last variable to be averaged. The following matrix illustrates the point:

					row means
	k	l	m	n	J_1
	o	p	q	r	J_2
	s	t	u	v	J_3
	w	x	y	z	J_4
column means	j_1	j_2	j_3	j_4	$\bar{\bar{g}}$ group mean

Table 1

Critical Sentences (Design 1): Mean Reaction Times of the Four Experimental Groups, Marginal Means, and the Grand Mean

		Type of Clause		
		Subordinate	Main	Totals
Probe Position	First Clause	761.18	745.74	753.46
	Second Clause	730.29	580.68	655.48
Totals:		745.73	663.21	704.47

Averaging the data points, or averaging row means, or averaging column means produces the same group mean:

$$\bar{g} = \frac{(k + 1 + \ldots + y + z)}{16} = \frac{J_1 + J_2 + J_3 + J_4}{4}$$

$$= \frac{j_1 + j_2 + j_3 + j_4}{4}$$

But, there is different variability associated with each method, since it is extremely unlikely that the variability of k,1, . . . y, z is equal to the variability of J_a's; and that either of these are equal to the variability of j_b's.

Accordingly, in the present study, the means of the treatment groups are examined in light of variability associated with the three methods of obtaining those means, the J_a row means are subject means; and a subjects analysis ($\underline{F}_i$) considers the between group differences relative to the differences among subjects. The j_b column means are sentence means; and a materials analysis (F_2) considers the between group differences relative to the differences among sentences. Clark's (1973) formula for min $\underline{F}'$ uses $\underline{F}_1$ and $\underline{F}_2$ to approximate the analysis computed on the raw data matrix.

Subject Analysis

Each subject's 17 responses to the critical sentences were combined into one mean reaction-time score (Table 2). This data was analyzed as a 2 x 2 between-subjects design, Probe Position (first or second clause) by Type of Clause (subordinate or main). The analysis of variance summary is

Table 2

Critical Sentences (Design 1): Mean Reaction Times of Each Subject across 17 Sentences

Subject	Subordinate Clause	Subject	Main Clause
	Probe Position: First Clause		
1	823.17	35	742.76
2	771.82	36	806.17
3	789.94	37	645.41
4	675.00	38	801.82
5	761.17	39	659.29
6	827.94	40	687.17
7	815.76	41	766.11
8	662.52	42	698.00
9	874.88	43	689.00
10	787.58	44	650.00
11	713.35	45	876.23
12	743.70	46	520.29
13	714.41	47	1038.94
14	730.64	48	636.58
15	662.35	49	940.35
16	667.00	50	564.35
17	918.76	51	955.11
	Probe Position: Second Clause		
18	932.82	52	633.25
19	692.11	53	524.11
20	655.17	54	560.23
21	697.17	55	474.29
22	661.17	56	700.82
23	764.82	57	569.64
24	710.41	58	635.29
25	789.29	59	531.17
26	645.58	60	584.94
27	567.05	61	695.47
28	755.70	62	447.88
29	633.11	63	544.17
30	426.47	64	542.82
31	809.70	65	453.11
32	717.05	66	596.35
33	602.70	67	687.52
34	854.52	68	680.35

presented in Table 3.

The main effect for Probe Position was significant ($F = 14.78$, $df = 1,64$, $p < .003$). For the critical sentences used, first-clause probes reliabily evoked longer reaction times than probes from the second clause. The main effect for Type of Clause was also highly significant ($F = 10.49$, $df = 1,64$, $p < .0019$). Reaction times to probes in a subordinate clause were longer than reaction times to main-clause probes. In addition, the interaction between Probe Position and Type of Clause was significant ($F = 6.93$, $df = 1,64$, $p < .01$). An inspection of the cell and marginal means presented in Table 1 indicates that, although the direction of each main effect is maintained at both levels of the other independent variable, the magnitude of that effect is not maintained.

An analysis of simple main effects was performed on the data (Table 4). The effect of Probe Position was significant only for probes in main clauses ($F = 20.98$, $df = 1,64$, $p < .003$); first-clause probes evoked longer reaction times than second-clause probes. The effect for Type of Clause was significant only for second-clause probes ($F = 17.23$, $df = 1,64$, $p < .0003$); subordinate-clause probes evoked longer reaction times than main-clause probes.

Language Materials Analysis

For each sentence, the responses of all subjects in each experimental group combined into one mean reaction-time

Table 3

Critical Sentences (Design 1): Analysis of Variance over Subjects (F_1)

Source	SS	df	MS	F	p
Probe Position (PP)	163,187.73	1	163,187.73	14.78	.0003
Type of Clause (TC)	115,769.42	1	115,769.42	10.49	.0019
PP x TC	76,509.76	1	76,509.76	6.93	.0106
Subjects within Groups	706,300.70	64	11,035.94		
Totals	1,061,767.61	67			

Table 4

Critical Sentences (Design 1): Analysis of Variance over Subjects (F_1)--Simple Main Effects

Source	SS	df	MS	F	p
Probe Position (PP) for Subordinate	8,109.70	1	8,109.70	.73	>.2
PP for Main	231,587.79	1	231,587.79	20.98	.0003
Type of Clause (TC) for First Clause	2,025.12	1	2,025.12	.18	>.2
TC for Second Clause	190,254.06	1	190.254.06	17.23	.0003
Subjects within Groups	706.300.70	64	11,035.94		

score (Table 5). This data was analyzed as a 2 x 2 design, Probe Position by Type of Clause; and, since the sampling variable (sentences) was identical for each cell, a completely repeated-measures analysis was performed. This was in contrast to the previous between-subjects design. The analysis of variance summary is presented in Table 6.

The pattern of significant results was identical to that found in the subjects analysis. Both main effects and the interaction were significant (Probe Position, $\underline{F}$ = 43.37, $\underline{df}$ = 1,16, $p < .0009$; Type of Clause, $\underline{F}$ = 41.37, $\underline{df}$ = 1,16, $p < .0009$; Probe Position by Type of Clause, $\underline{F}$ = 15,20, $\underline{df}$ = 1,16, $p < .01$).

An analysis of simple main effects was performed on the data (Table 7). The Probe-Position effect was not significant for subordinate-clause probes ($\underline{F}$ = 2.20, $\underline{df}$ = 1,16). For probes in the main clause, the Probe-Position effect was significant ($\underline{F}$ = 45.23, $\underline{df}$ = 1,16, $p < .0009$): first-clause probes were longer than second-clause probes. The Type-of-Clause effect was significant only for second-clause probes ($\underline{F}$ = 69.44, $\underline{df}$ = 1,16, $p < .0009$): subordinate-clause probes were longer than main-clause probes.

Comparison of Subjects and Language Materials Analysis

In all $\underline{F}_1$ - $\underline{F}_2$ comparisons, the variation due to treatment effects (and in this two-way analysis, the variation due to the interaction of treatment effects) will be identical

Table 5

Critical Sentences (Design 1): Mean Reaction Times to Each Sentence across 17 Subjects

Sentence	Subordinate Clause	Sentence	Main Clause
	Probe Position: First Clause		
1	818.47	1	793.64
2	851.00	2	846.35
3	803.52	3	791.00
4	835.82	4	784.52
5	814.35	5	837.05
6	716.84	6	698.70
7	737.11	7	682 35
8	744.25	8	763.23
9	720.58	9	901.52
10	684.05	10	669.82
11	802.41	11	775.41
12	766.17	12	619.94
13	790.04	13	735.82
14	770.70	14	705.11
15	701.88	15	694.11
16	741.76	16	683.35
17	590.23	17	695.64
	Probe Position: Second Clause		
1	707.00	1	655.47
2	782.17	2	663.47
3	767.05	3	629.88
4	691.70	4	536.23
5	675.35	5	642.58
6	645.82	6	508.82
7	631.58	7	490.52
8	764.58	8	565.58
9	694.58	9	550.88
10	728.23	10	553.58
11	819.47	11	560.82
12	752.58	12	521.47
13	788.52	13	544.00
14	756.29	14	606.17
15	667.29	15	603.11
16	801.35	16	663.00
17	740.29	17	564.94

Table 6

Critical Sentences (Design 1): Analysis of Variance over Language Materials (F_2)

Source	SS	df	MS	F	p
Probe Position (PP)	163,187.73	1	163,187.73	43.37	.00009
Sentences	161,344.15	16	10,084.00		
PP x Sentences	60,198.25	16	3,762.39		
Type of Clause (TC)	115,769.42	1	115,769.42	41.37	.00009
TC x Sentences	44,768.11	16	2,798.00		
PP x TC	76,709.76	1	76,709.76	15.20	.01
PP x TC x Sentences	80,720.36	16	5,045.02		
Totals	702,697.78	67			

Table 7

Critical Sentences (Design 1): Analysis of Variance over Language Materials (F_2)-- Simple Main Effects

Source	SS	df	MS	F	p
Probe Position (PP) for Subordinate	8,109.70	1	8,109.70	2.20	.2
PP at Subordinate x Sentences	59,003.61	16	3,687.72		
PP for Main	231,587.79	1	231,587.79	45.23	.00009
PP at Main x Sentences	81,915.20	16	5,119.70		
Type of Clause (TC) for First Clause	2,025.12	1	2,025.12	.39	.2
TC_1 x Sentences	81,657.56	16	5,103.59		
TC for Second Clause	190,254.06	1	190.254.06	69.44	.00009
TC_2 x Sentences	43,831.47	16	2,739.46		

because this variation is based in means, and as the example matrix earlier illustrated, the same mean value results from either method of calculation. But the variation associated with error terms in each design, and hence the total variation of each design, will be different. This is because both error and total variation terms take into account the elements that contribute to the means, and in $\underline{F}_1$ and $\underline{F}_2$ analyses, those elements are different. For design 1, compare Tables 2 and 5. Table 2 (subject means) lists the elements that contribute to the $\underline{F}_2$ analysis. Although the means for the data in Tables 2 and 5 are identical, the variability of the numbers that contribute to those means is quite different. The $\underline{F}_1$'s and $\underline{F}_2$'s of designs 2 and 3 may be similarly compared. In all cases, variations computed from means will be alike; variations computed from elements will differ.

Analysis over Subjects and Language Materials

Significant $\underline{F}_1$ and $\underline{F}_2$ analyses imply that results are presumed to be reliable when materials are held constant and subjects are changed ($\underline{F}_1$), and when subjects are held constant and materials are changed ($\underline{F}_2$). In order to examine the statistical evidence for the reliability of the effects when materials and subjects are simultaneously changed, an analysis in which both subjects and materials are sampling variables (i.e., random variables) needs to be considered. But both Winer (1971) and Clark (1973) point out that, for designs with

fixed variables and at least two random variables, there is no single error term that provides an appropriate test of fixed factor main effects. That is, the expected mean square for the treatment variable (a fixed factor) contains variance components due to the treatment effect and error, and variance components due to the interaction of the fixed factor with each random variable, both separately and jointly. Using the mean square for the interaction of the fixed factor with one of the random variables as the error term is inadequate since the variance component due to the interaction of the fixed factor with the other random variable still remains in the expected mean square for the treatment variable. Table 8, adapted from Winer (1971, p. 572) and Clark (1973, p. 344) illustrates the problem for a repeated-measures design when the researcher wishes to test the treatment effect against both subjects and materials simultaneously. An $\underline{F}_1$ analysis is inadequate since,

$$\frac{\text{Treatments}}{\text{Treatments X Subjects}} = \frac{\sigma_\varepsilon^2 + \sigma_{tms}^2 + q\sigma_{ts}^2 + r\sigma_{tm}^2 + qr\sigma_t^2}{\sigma_\varepsilon^2 + \sigma_{tms}^2 + q\sigma_{ts}^2}$$

$$= r\sigma_{tm}^2 + qr\sigma_t^2$$

Similarly an $\underline{F}_2$ analysis is inadequate since,

$$\frac{\text{Treatments}}{\text{Treatments X Materials}} = \frac{\sigma_\varepsilon^2 + \sigma_{tms}^2 + q\sigma_{ts}^2 + r\sigma_{tm}^2 + qr\sigma_t^2}{\sigma_\varepsilon^2 + \sigma_{tms}^2 + r\sigma_{tm}^2}$$

$$= q\sigma_{ts}^2 + qr\sigma_t^2$$

Table 8

Sources of Variance and Expected Mean Squares for a Repeated-Measures Design with One Fixed Factor and Two Random Factors

Source	Expected Value of Mean Square
T:Treatments (p)	$\sigma^2 + \sigma^2_{tms} + q\sigma^2_{ts} + r\sigma^2_{tm} + qr\sigma^2_{t}$
M:Language Materials (q)	$\sigma^2 + p\sigma^2_{ms} + pr\sigma^2_{m}$
S:Subjects (r)	$\sigma^2 + p\sigma^2_{ms} + pq\sigma^2_{a}$
TxM:Treatments x Language Materials	$\sigma^2 + \sigma^2_{tms} + r\sigma^2_{tm}$
TxS:Treatments x Subjects	$\sigma^2 + \sigma^2_{tms} + q\sigma^2_{ts}$
MxS:Language Materials x Subjects	$\sigma^2 + p\sigma^2_{ms}$
TxMxS:Treatments x Language Materials x Subjects	$\sigma^2 + \sigma^2_{tms}$

In order to obtain a ratio that yields only the variance component due to treatments, certain mean squares from Table 8 need to be combined (by simple arithematic operations) to produce composite mean squares. Two possibilities result:

(1) $$\frac{MS_{Treatments}}{MS_{TxS} + MS_{TxM} - MS_{TxMxS}}$$

$$= \frac{\sigma_\epsilon^2 + \sigma^2_{tms} + q\sigma^2_{ts} + r\sigma^2_{tm} + qr\sigma^2_t}{\sigma_\epsilon^2 + \sigma^2_{tms} + q\sigma^2_{ts} + \sigma_\epsilon^2 + \sigma^2_{tms} + r\sigma^2_{tm} - (\sigma_\epsilon^2 + \sigma^2_{tm})}$$

$$= \frac{\sigma_\epsilon^2 + \sigma^2_{tms} + q\sigma^2_{ts} + r\sigma^2_{tm} + qr\sigma^2_t}{\sigma_\epsilon^2 + \sigma^2_{tms} + q\sigma^2_{ts} + r\sigma^2_{tm}}$$

$$= qr\sigma^2_t$$

(2) $$\frac{MS_{Treatments} + MS_{TxMxS}}{MS_{TxS} + MS_{TxM}}$$

$$= \frac{\sigma_\epsilon^2 + \sigma^2_{tms} + q\sigma^2_{ts} + r\sigma^2_{tm} + qr\sigma^2_t + \sigma_\epsilon^2 + \sigma^2_{tms}}{\sigma_\epsilon^2 + \sigma^2_{tms} + q\sigma^2_{ts} + \sigma_\epsilon^2 + \sigma^2_{tms} + r\sigma^2_{tm}}$$

$$= qr\sigma^2_t$$

Winer recommends (2) since (1) calls for a subtraction that could possibly lead to a negative denominator. Since composite mean squares are used, the ratios are called quasi-$\underline{F}$'s.

Clark (1973) concurs with Winer's preference for (2) and labels that quasi-$\underline{F}$ ratio $\underline{F}'$. In addition, Clark suggests that a combination of the simpler $\underline{F}_1$ and $\underline{F}_2$ analyses can be

used to approximate $\underline{F}'$. He argues as follows:

$$\underline{F}' = \frac{MS_T + MS_{TxMxS}}{MS_{TxS} + MS_{TxM}}$$

MS_{TxMxS} is the only term of the four that requires the use of the complete unaveraged data matrix. If the minimum value of MS_{TxMxS} (zero) is used, the term can be deleted and,

$$\min \underline{F}' = \frac{MS_T}{MS_{TxS} + MS_{TxM}}$$

In terms of expected mean squares,

$$\min \underline{F}' = \frac{qr\sigma_t^2}{\sigma_\epsilon^2 + \sigma_{tms}^2}$$

Obviously, $\dfrac{qr\sigma_t^2}{\sigma_\epsilon^2 + \sigma_{tms}^2}$

is always less than or equal to $qr\sigma_t^2$. Thus, whenever min $\underline{F}'$ is significant, $\underline{F}'$ is significant. Clark also presents a computationally simple formula for computing min $\underline{F}'$:

$$\min \underline{F}'_{(i,j)} = \frac{F_1 F_2}{F_1 + F_2}$$

i equal the degrees of freedom for the treatment mean square, and j is the nearest integer that results from the following formula:

$$j = \frac{(F_1 + F_2)^2}{(\frac{F_1^2}{n_2} + \frac{F_2^2}{n_1})}$$

n_1 equals the degrees of freedom for the $\underline{F}_1$ error term and n_2 equals the degrees of freedom for the $\underline{F}_2$ error term.

Min $\underline{F}'$ values were calculated for all main and simple effects in which both $\underline{F}_1$ and $\underline{F}_2$ or both were significant. There were no main or simple main effects in which one but not both of the $\underline{F}_1$ and $\underline{F}_2$ analyses were significant. Table 9 lists the results of the min $\underline{F}'$ calculations. The pattern of results is identical to the $\underline{F}_1$ and $\underline{F}_2$ patterns. For all effects (main and simple main) where $\underline{F}_1$ and $\underline{F}_2$ were significant, min $\underline{F}'$ was significant. Table 10 presents a summary of the $\underline{F}_1$, $\underline{F}_2$ and min $\underline{F}'$ findings for the critical sentences. Figure 6 is a graphic representation of the results of the statistical analyses. If groups did not differ significantly from one another, then the points on the graph representing those groups were enclosed in a common box.[31]

Analysis of Filler Sentences (Design 2)

The data from the filler sentences for each subject group was combined into one mean reaction-time score. The mean of each subject group as well as the grand mean of all four subject groups are presented in Table 11.

---

[31] In this Figure, and in Figures 7, 8, and 10, the abcissa is nominal. Therefore, points are connected by a line simply for visual clarity in identifying common levels of an independent variable. There is no attempt to imply any linear relationship between connected levels. Line graphs are used instead of bar graphs because one figure (10) represents a simultaneous look at the three designs. The overlapping nature of this illustration precludes the use of a bar graph.

Table 9

Critical Sentences (Design 1):

Min $\underline{F}'$ Statistics

Main Effects	min $\underline{F}'$	df	p
Source:			
Probe Position (PP)	11.02	1, 79	.001
Type of Probe (TC)	8.36	1, 80	.01
PP x TC	4.76	1, 74	.05
Simple Main Effects			
Source:			
PP at Subordinate	*	*	*
PP at Main	14.33	1, 74	.001
TC at First Clause	*	*	*
TC at Second Clause	13.80	1, 80	.001

*not calculated (both $\underline{F}_1$ and $\underline{F}_2$ nonsignificant)

Table 10

Critical Sentences (Design 1): Summary of Results from F_1, F_2, and min F' Analyses

Main Effects	F_1	F_2	min F'
Source:			
Probe Position (PP)	ALL SIGNIFICANT ($p < .05$)		
Type of Clause (TC)			
PP x TC			
Simple Main Effects			
Source:			
PP at Subordinate	all nonsignificant		
PP at Main	ALL SIGNIFICANT ($p < .05$)		
TC at First Clause	all nonsignificant		
TC at Second Clause	ALL SIGNIFICANT ($p < .05$)		

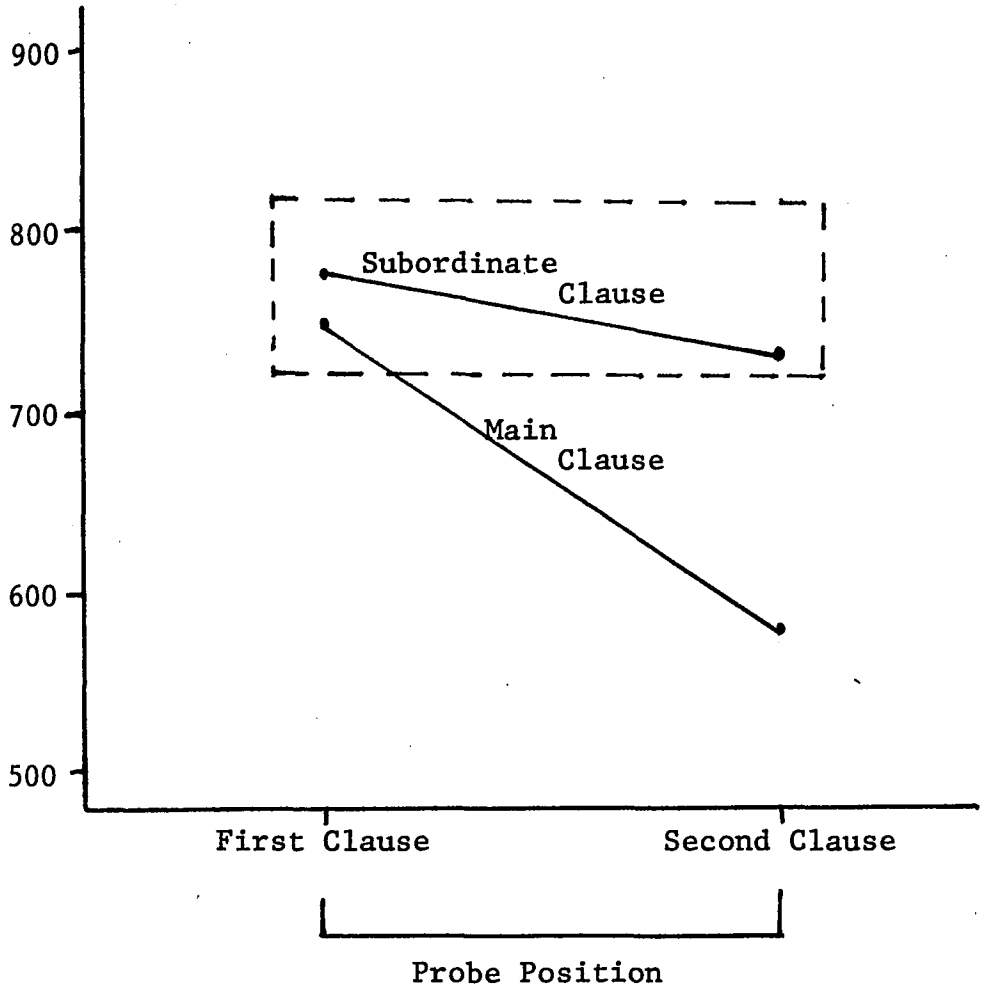


Figure 6. The four experimental groups of design 1 (critical sentences).

Table 11

Filler Sentences (Design 2): Mean Reaction Times for Each Subject Group and the Grand Mean

	Subject Groups			
Grand Mean	1	2	3	4
662.25	633.23	684.00	666.97	664.81

Subjects Analysis

Each subject's 17 responses to the filler sentences were combined into one mean reaction-time score (Table 12). This data was analyzed as a between-subjects, one-factor design. The analysis of variance summary is presented in Table 13. The differences among the groups were not significant.

Language Materials Analysis

For each sentence, the responses of all subjects in each group were combined into one mean reaction-time score (Table 14). A one-way analysis of variance was computed and, with sentences as the sampling variable, a repeated-measures analysis was appropriate since, in each cell, the same sentences were used. This was in contrast to the previous between-subjects design. Table 15 is the summary for the analysis of variance. Again, the differences among the groups were not significant.

Analysis over Subjects and Language Materials

Since both the subjects analysis ($\underline{F}_1$) and the materials analysis ($\underline{F}_2$) were nonsignificant, the analysis considering subjects and materials as sampling variables (min $\underline{F}'$) was not calculated. Figure 7 is a graphic representation of the results of the $\underline{F}_1$ and $\underline{F}_2$ analyses. Since neither analysis showed a difference between the groups, the points that represent each group are enclosed in a common box.

Table 12

Filler Sentences (Design 2): Mean Reaction Times

For Each Subject across 17 Sentences

Subject	Group 1	Subject	Group 2
1	631.29	18	864.94
2	611.94	19	695.52
3	645.35	20	580.23
4	540.76	21	666.05
5	663.70	22	726.35
6	648.05	23	769.70
7	696.70	24	690.82
8	582.00	25	749.70
9	706.35	26	595.88
10	643.35	27	481.88
11	503.52	28	704.58
12	652.17	29	688.11
13	594.64	30	864.52
14	735.58	31	657.29
15	631.11	32	597.52
16	564.00	33	598.82
17	715.64	34	711.82
	Group 3		Group 4
35	772.35	52	717.11
36	817.64	53	620.58
37	635.82	54	637.47
38	824.41	55	615.23
39	684.52	56	731.70
40	665.70	57	641.94
41	597.17	58	674.76
42	575.35	59	647.58
43	575.23	60	636.00
44	664.41	61	700.41
45	699.82	62	564.35
46	585.70	63	622.64
47	767.17	64	736.05
48	564.17	65	635.88
49	676.22	66	702.05
50	516.76	67	706.35
51	716.82	68	662.52

Table 13

Filler Sentences (Design 2): Analysis of Variance over Subjects (F_1)

Source	SS	df	MS	F	p
Groups	22,850.24	3	7,616.74	1.26	>.2
Subjects within Groups	385,041.00	64	6,016.26		
Total	407,891.24				

Table 14

Filler Sentences (Design 2): Mean Reaction Times to Each Sentence across 17 Subjects

	Subject Groups			
Sentence	1	2	3	4
1	855.88	896.70	800.47	790.82
2	805.35	735.40	817.24	921.47
3	687.64	689.99	798.59	760.76
4	514.58	632.11	578.06	608.00
5	713.70	691.70	724.89	708.76
6	609.35	705.52	619.83	593.76
7	568.17	632.34	655.24	568.41
8	676.82	599.17	559.83	549.23
9	439.52	633.22	568.71	591.11
10	464.41	618.64	772.41	601.35
11	599.05	652.97	586.71	586.70
12	722.17	719.11	719.53	847.94
13	641.35	739.40	652.00	601.58
14	553.94	731.70	688.71	625.88
15	528.88	621.87	554.59	562.17
16	701.05	662.34	538.77	716.58
17	634.29	666.52	703.83	654.11

Table 15

Filler Sentences (Design 2): Analysis of Variance over Language Materials (F_2)

Source	SS	df	MS	F	p
Between Subjects	493,887.36	16	30,867.93		
Groups	22,850.24	3	7,616.74	2.53	>.2
Groups x Sentences	14,400.00	48	3,000.00		
Total	660,745.60				

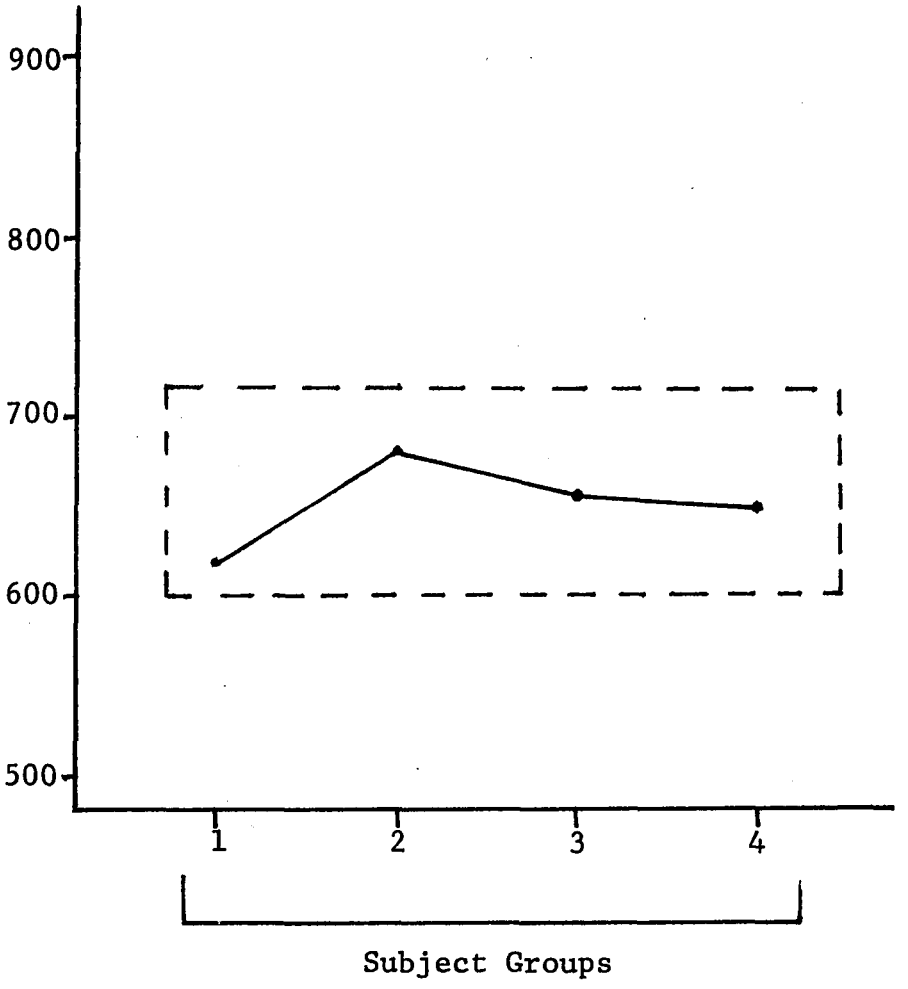


Figure 7. The four subject groups of design 2 (filler sentences).

Analysis of Probe Types (Design 3)

Each subject group heard the same sentences, but different probe words for those sentences. Only one type of probe (IN, OUT, RHYME, ADJ-IN), was presented to each group. The data from each probe-type group was combined into one mean reaction-time score. The mean for each group as well as the grand mean of the four groups is presented in Table 16.

Subjects Analysis

Each subject's 17 responses to the filler sentences of design 3 were combined into one mean reaction-time score (Table 17). This data was analyzed as a between-subjects, one-factor design. The analysis of variance summary is presented in Table 18. The treatment effect of Type of Probe was significant ($F = 17.93$, $df = 3,64$, $p < .00009$).

A Newman-Keuls test was performed to determine pairwise differences between group means. The difference between the IN and ADJ-IN groups was not significant. All other pairwise combinations were significantly different ($p < .01$).

Language Materials Analysis

For each sentence, the responses of all subjects in each probe-type group were combined into one mean reaction-time score (Table 19). This data was analyzed as a 1 x 4 design; and since the sampling variable (sentences) was identical for each cell, a one-way repeated-measures analysis was performed. This was in contrast to the previous between-

Table 16

Probe Types (Design 3): Mean Reaction Time of the Four Experimental Groups and the Grand Mean

	Probe Types			
Grand Mean	IN	OUT	RHYME	ADJ-IN
730.36	640.80	755.67	862.30	662.67

Table 17

Probe Types (Design 3): Mean Reaction Times

for Subjects across 17 Sentences

Subject	IN	Subject	OUT
1	661.58	18	877.17
2	649.70	19	740.76
3	716.41	20	589.41
4	475.47	21	713.35
5	662.35	22	711.47
6	672.52	23	890.58
7	760.82	24	751.47
8	590.88	25	914.23
9	732.05	26	690.29
10	707.52	27	527.47
11	536.88	28	753.52
12	595.94	29	657.70
13	633.70	30	862.47
14	622.47	31	834.11
15	627.70	32	765.11
16	561.05	33	729.47
17	674.23	34	837.88
	RHYME		ADJ-IN
35	872.64	52	711.64
36	974.29	53	599.64
37	732.25	54	628.44
38	886.05	55	652.17
39	882.05	56	797.29
40	1035.41	57	676.64
41	837.05	58	712.17
42	923.17	59	718.58
43	792.47	60	627.76
44	770.29	61	816.35
45	945.35	62	440.41
46	631.17	63	663.17
47	1016.70	64	647.23
48	772.17	65	467.41
49	1006.23	66	635.23
50	759.41	67	736.82
51	827.82	68	739.52

Table 18

Probe Types (Design 3): Analysis of Variance over Subjects (F_1)

Source	SS	df	MS	F	p
Type of Probe	524,765.87	3	174,921.95	17.93	.00009
Subjects within Groups	624,703.05	64	9,760.98		
Totals	1,149,468.92	67			

Table 19

Probe Types (Design 3): Mean Reaction Times

to Each Sentence across 17 Subjects

Sentence	IN	OUT	RHYME	ADJ-IN
1	806.52	867.70	1788.70	861.05
2	848.82	794.24	1004.47	796.88
3	682.35	723.94	967.88	751.88
4	656.00	698.11	913.82	606.00
5	723.70	761.05	726.64	735.64
6	677.05	736.94	723.35	671.47
7	575.35	641.64	823.23	548.41
8	555.47	680.94	815.00	884.00
9	590.41	777.00	672.05	775.35
10	562.88	842.11	799.47	546.23
11	640.11	644.00	707.05	866.35
12	515.64	809.41	762.05	578.17
13	513.88	794.58	842.52	431.05
14	695.88	761.29	652.82	598.00
15	489.76	648.82	736.70	630.35
16	692.70	904.00	882.64	465.23
17	654.76	760.64	846.23	519.29

subjects analysis. The analysis of variance summary is presented in Table 20. The treatment effect of Type of Probe was significant ($\underline{F} = 8.96$, $\underline{df} = 3,48$, $p < .05$).

A Newman-Keuls test was performed to determine pair-wise differences between group means. The pattern of results is identical to that found in the subjects analysis: the difference between the IN and ADJ-IN groups was not significant; all other pair-wise combinations were significantly different ($p < .01$).

The standard deviations for the sentence means for each group were computed (Table 21). From smallest to largest, the groups ordering of the standard deviations is: OUT, IN, ADJ-IN, RHYME.

Analysis over Subjects and Language Materials

Because both $\underline{F}_1$ and $\underline{F}_2$ were significant, a min $\underline{F}'$ was calculated. Again, the treatment effect of Type of Probe was significant (min $\underline{F}' = 5.97$, $\underline{df} = 3,91$, $p < .001$). Figure 8 is a graphic representation of the results of the statistical analyses.

Table 20

Probe Types (Design 3): Analysis of Variance over Language Materials (F_2)

Source	SS	df	MS	F	p
Between Sentences	718,313.03	16	44,894.56		
Type of Probe	524,765.87	3	174,921.95	8.96	.05
Type of Probe x Sentences	936.340.06	48	19,517.50		
Totals	2,179,918.96	67			

Table 21

Probe Types (Design 3): Standard Deviations of Sentence Means around the Probe-Type Means

Type of Probe			
IN	OUT	RHYME	ADJ-IN
100.97	77.29	256.44	146.47

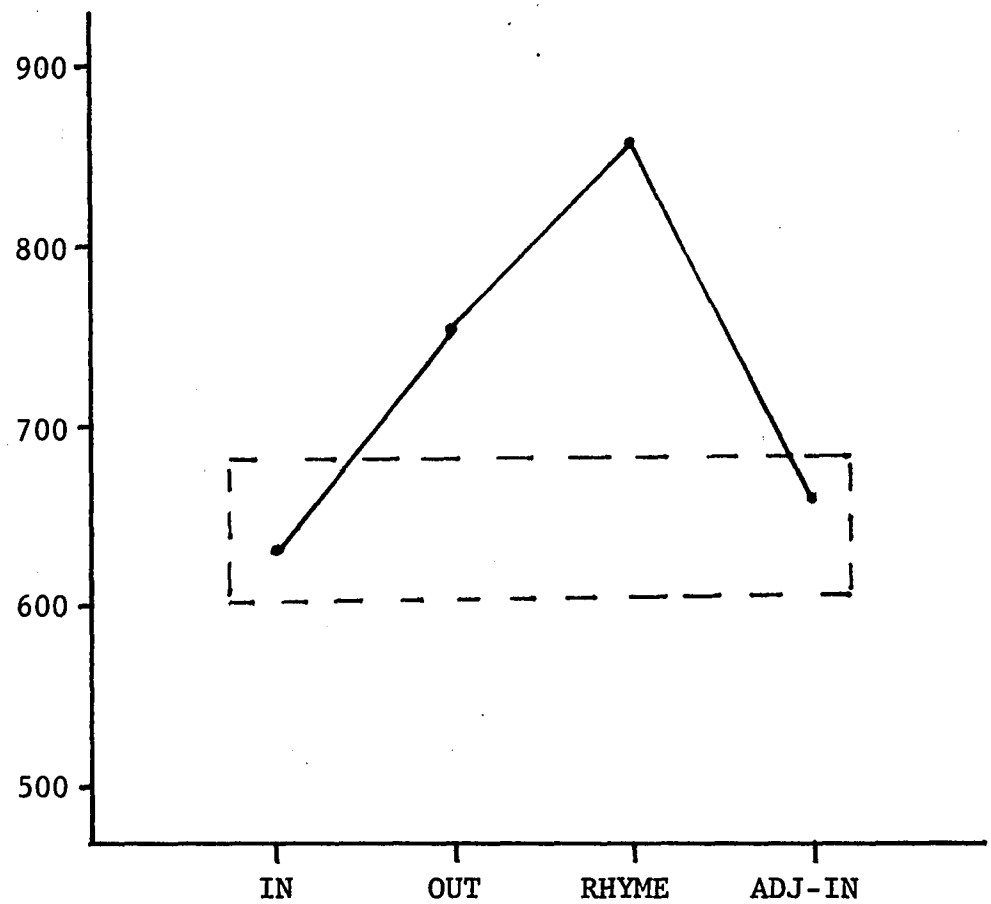


Figure 8. The four experimental groups of design 3 (probe types).

DISCUSSION

Experimental Psycholinguistics: Clausal, Lexical and Surface-Structure Analyses (Design 1)

The significant results of the analyses of the critical sentences (design 1) can be interpreted as consequences of the treatment variables. The pattern of the $\underline{F}_1$, $\underline{F}_2$ and min $\underline{F}'$ analyses is consistent (Table 10). Effects were either always significant or always insignificant across the three analyses. Thus, the main and interactive effects of Probe Position and Type of Clause are reliable over both subject and material populations.

It is necessary, however, to designate as specifically as possible, the characteristics of populations to which the results of this study apply. The subject population is defined as complex English sentences having one adverbial subordinate clause. It is beyond the scope of this research to predict the effects of the variables studied on other types of subordinate clauses. Although only complex sentences with adverbial clauses are considered, the lexical items of each sentence readily allow for the construction of complement clause versions of each sentence. The population contains both clause orders: subordinate-main and main-subordinate. All subordinate clauses are introduced by an adverb--a clear surface-structure marking of the subordination. Moreover, in

the sample, all sentences were 24 syllables long. Although it would be quite meaningless to limit the findings of this research to only those sentences that are 24 syllables long, Lang (1974) has suggested that within the probe paradigm, reaction-time predictions about sentences can be made only if the syllable lengths of sentences to be compared differ by no more than 4 syllables. The population under consideration contains groups of sentences to be compared that conform to this restriction. Finally, the probe task is an extremely easy task for any native speaker. Decisions about the inclusion of the probe word in the previous sentence are generally reported within three-quarters of a second. Thus, it is a very automatic level of linguistic processing that is reflected in the dependent measures.

Both Probe Position and Type of Clause produced significant effects (Tables 3, 6, 9 and 10), and those effects were in the predicted directions. The present results agree with the findings of Caplan (1972), Kornfeld (1974) and Lang (1974), that the mean reaction time for first-clause probes is significantly longer than the mean reaction time for second-clause probes. Similarly, results support Kornfeld's (1974) suggestion that reaction times to probes from subordinate clauses are on the average longer than reaction times to probes from main clauses. Both of these findings are consistent with the theoretical framework that has been presented and which is summarized as follows:

1. Information about sentences is stored clause by clause and immediate accessibility of clausal material is inversely related to the temporal order of the clause: information stored in earlier clauses is less accessible than information from more recently stored clauses. The segmentation of the sentence into clausal units is presumably accomplished by the repeated application of heuristic A (see Introduction, p. 22):

   A: Segment together any sequence X . . . Y, in which the members could be related by primary internal structural relations, "actor-action object . . . modifier" (Bever, 1970, p. 290).

2. The primary content of sentences is conveyed by the main clause; supplementary information is included in subordinate clauses. Results suggest that clausal material is stored in a hierarchial fashion such that primary content is more accessible than supplementary information. After a clausal segmentation has been accomplished by the repeated application of heuristic A, then the relationship between those clauses can be determined by an application of heuristic B (see Introduction, p. 32).

   B: Take the verb which immediately follows the initial noun of a sentence as the main verb unless there is a surface structure mark of an embedding (Fodor et al., (1974, p. 356).

For all the sentences of this design, the application of heuristic B provides a correct description of the hierarchical

structure between the clauses: whenever the first verb of the sentence is not the main verb, that clause is introduced by an adverb that is a clear surface-structure mark of the subordination.

In addition to expecting main effects due to both Probe Position and Type of Clause, an additive (or noninteractive) relationship between the two variables was predicted. The basis of the prediction was its simplicity--if each variable is effective, then the combination of the two variables together might be additively strong. The only previous research to simultaneously examine the effects of Probe Position and Type of Clause (Kornfeld, 1974), found the relationship between the two variables was unstable across various experiments. It is possible that the instability was due to changes in the perceptual complexity of the lexical items used in the various experiments. One experiment examined only complex sentences with adverbial clauses where the main verb did not allow complement constructions. Another experiment studied complex sentences with adverbial, relative, and complement clauses. Thus, since lexical items were remarkably stable across the various sentence versions, sentences with adverbial clauses frequently contained lexical markers that could dominate a complement construction.

The predicted additivity did not occur. $\underline{F}_1$, $\underline{F}_2$ and min $\underline{F}'$ analyses showed a significant Probe Position by Type of Clause interaction (Tables 3, 6, 9 and 10). Figure 9 describes the interactive relationships between the experi-

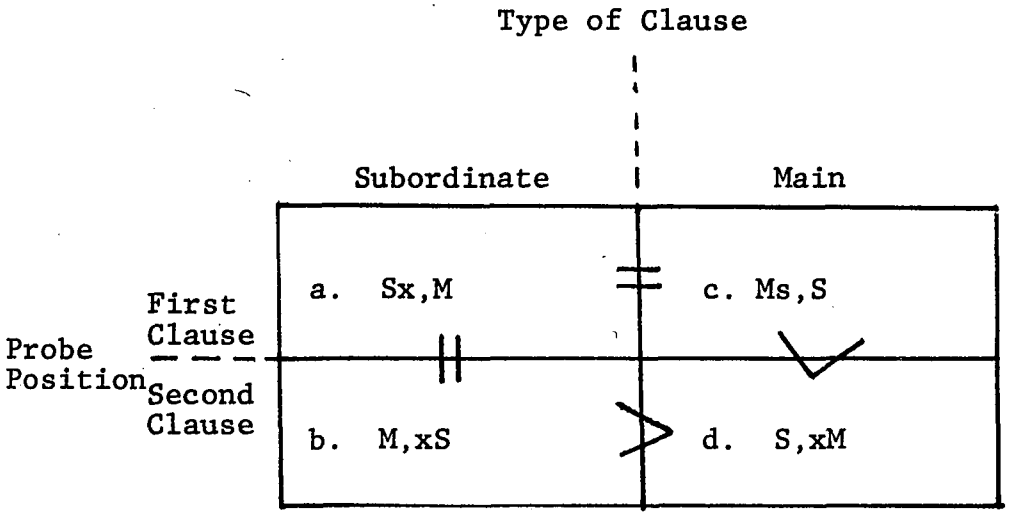


Figure 9. Comparison of the differences between the experimental groups of design 1 (critical sentences).

mental cells that were revealed by simple main effects analyses (see also Tables 4, 7, 9, and 10 and Figure 6). Three of the cells do not differ significantly from each other, and each of the three is significantly larger than the fourth cell. Neale and Liebert (1973) have called such an interaction "terminative":

> A terminative interaction is one in which two or more variables are clearly effective in modifying behavior, but, when combined, their effect is not increased over what either would do alone (p. 65).

For the present data, a qualifying amendment will be added to the definition: A terminative interaction is one in which two variables are clearly effective in modifying behavior, under specifiable conditions, but, when combined, their effect is not increased over what either would do alone.

The Probe Position variable differentially effects reaction times only when probes are from main clauses (cell c.--Mx,S--is greater than cell d.--S,xM--and cell a.--Sx,M--is equal to cell b.--M,xS); the Type of Clause variable differentially affects reaction times only when probes are from second clauses (cell b.--M,xS--is greater than cell d.--S,xM--and cell a.--Sx,M--is equal to cell c.--Ms,S--). These effects can be theoretically summarized as follows: the effect of Probe Position and Type of Clause can be seen only when that variable operates on the level of the other variable that is most accessible to immediate memory. Neither variable is effective at the less accessible level of the other independent variable.

These results imply that sentential material from complex sentences with adverbial clauses is retrieved by a process that shows a ceiling effect after one increment of complexity. Cell d.--S,xM--involves the simplest task: recognizing information from a clause that is both most recent and most important. Cells b. and c. each add one increment of complexity to that task: recognition of information from a less important but still most recent clause (M,xS) or recognition of information from the less recent but still most important clause (Mx,S). An inspection of the analyses of simple main effects (Tables 4, 7, and 9) suggests that the increments of complexity added by cell b. or c. are approximately equal. Moreover, since cell a.--Sx,M--is not different from either cell b. or c., then the implication is that the task associated with a. is not more complex than the tasks associated with b. or c. The reaction-time increment from two sources--probe in the first clause, probe in a subordinate clause--is not any greater than the reaction-time increment from either source considered separately.

One reason for this pronounced ceiling effect might be that many of the sentences included syntactically complex lexical markers--markers that can dominate a complement construction. As noted in the introductory section on Lexical Analysis (pp. 25-28), such items may add perceptual complexity to a sentence. Thus there may be a consistent source of complexity present in the experiment. It is uncertain whether this complexity interacted with the experimental

results. A replication of this study with syntactically simpler lexical items is thus necessary. If such a study reproduces the results of the present study, then it would appear that clausal information that is less accessible (either because of its temporal order, or because of its subordinate construction, or both) is held in a kind of undifferentiated storage. That is, material from a clause which is both recent and main is placed in one category; material from all other clauses is placed in another less accessible category. There are different reasons why material may be stored in that second category, yet retrieval of any information from that category involves a similar process. If, on the other hand, results from a study with syntactically simpler lexical items yield an additive relationship between the Probe Position and Type of Clause variables, then a more hierarchical storage and retrieval model would be postulated. That is, again, recent main-clause information would be the most accessible; recent subordinate-clause information and less recent main-clause information would be in the next accessible category; and less recent subordinate-clause information would be least accessible. Should this occur, a comparison of the studies with simple and complex lexical items would imply that the inclusion of more complex lexical items in a sentence inhibits the discriminatory power of the storage and retrieval process as it applies to clausal material.

In sum, this study lends further support to the already widely held contention that listeners are sensitive to clausal analyses when they hear sentences. The storage and retrieval of information from each clause is influenced by the temporal order of the clause and by the main or subordinate function served by the clause in the sentence. In addition, the lexical complexity of the words in the sentence may determine the degree to which certain combinations of those temporal and structural variables are differentiated in the storage and retrieval processes.

Experimental Psycholinguistics: Methodology (Design 2)

In the previous section, the results of design 1--the mean reaction times for the four groups--were discussed as effects of the manipulated independent variables of Probe Position and Type of Clause. This causal link between the independent variables and the dependent variable can be inferred when the researcher is relatively confident about the internal validity of the study: that is, that the comparison groups differed only with respect to the independent variables under investigation. Of course, in an absolute sense, no experiment is ever truly internally valid--one can always think of some differences between the groups other than those associated with the independent variables. The question is, what is the nature of those differences? Are they theoretically relevant, systematic, and/or avoidable? When those

differences are brought up, what arguments are there to mitigate their influence on the dependent variable?

Recall that the inclusion of the filler sentences added to the external validity of the study. The analysis of data from those filler sentences (design 2) can add methodological strength to the interpretation of the results in design 1. Results from the filler analyses can contribute to the internal validity of the experimental study. In particular, for the present study, the filler analyses addressed the following questions:

1. Were there any initial differences between the groups? If there were, then those group differences were confounded with the treatment differences in design 1.
2. Were there any set differences created by the addition of design 3 (probe types) to the study? Presumably, if the data from design 1 were influenced by a response set to some instances of a particular type of probe, then the data from design 2 were also equally confounded. Again, differences between the groups in design 2 would suggest that some type of confounding was present in design 1, and so, causal inferences about the relationship between the independent and dependent variables would be suspect.

For the present study it was found that when the four subject groups were treated alike, none differed significantly from one another (Tables 13 and 15 and Figure 7). The mean differences between the groups were tested against the differences among subjects within a group (an F_1 analysis) and

against the differences among the interactions of the groups with the sentences (an F_2 analysis). Thus, the assumption of the equivalence of subject groups was supported; or, alternatively, any context differences that resulted from the confounding of designs 1 and 3 was not large enough to infer that different subject groups formed different response sets. The results from design 1 are thus more confidently attributed to the effects of the independent variables of Probe Position and Type of Clause.

Of course, one must be cautious in infering theoretical support from nonsignificant results. Strictly speaking, within the Fisher model of analysis of variance, one is not allowed to "accept" the null hypothesis; one either rejects or fails to reject it. The main reason for not directly accepting the null hypothesis is that groups can be nonsignificantly different because in fact, the groups are alike, or because the statistical test lacks power (too few subjects, an insensitive dependent variable, etc.). In single analyses, there is no way to choose between these alternatives.

Consider, however, the schematic arrangement of the present study: there are three distinct but related designs. By inspecting the details of each design and comparing their shared features, a statement regarding relative power of each design can be made. Although each design has different independent variables, the subjects, the task, and the response measure for all designs are identical; the same number of sentences are used and all sentences in the design contain

the same number of syllables. Moreover, the sentences from all three designs were combined, in a random order, into one tape. Thus, although the data collected are separated into three separate analyses, the collection process itself involved only one experimental session. There is thus no need to worry about sequential or carry-over effects that typically might be a problem if the same subjects are used in three distinct experiments. Designs 2 (subject groups) and 3 (probe types) have comparable statistical power and sensitivity. Both are one-way analyses of variance; both $\underline{F}_1$'s are between-subjects analyses; both $\underline{F}_2$'s are repeated measures. Design 1 (critical sentences) is a two-by-two design.

By comparing the pattern of results in the three designs, the nonsignificant results from the filler analyses (design 2) can be interpreted. In Figure 10, the means of the four groups are graphed separately for each design. One can readily see that the relationship among the four groups differs with each design. There is not a common overriding group pattern independent of the treatment conditions of each design. Moreover, since both the critical sentences of design 1 and the probe types of design 3 showed significant effects, it is quite unreasonable to suppose that the nonsignificant results from the filler sentences are due to a lack of power or sensitivity. Rather, the implication is that the non-significance represents a statistical judgment regarding the equivalence of the groups over subjects and sentences; and the lack of a context effect across the groups.

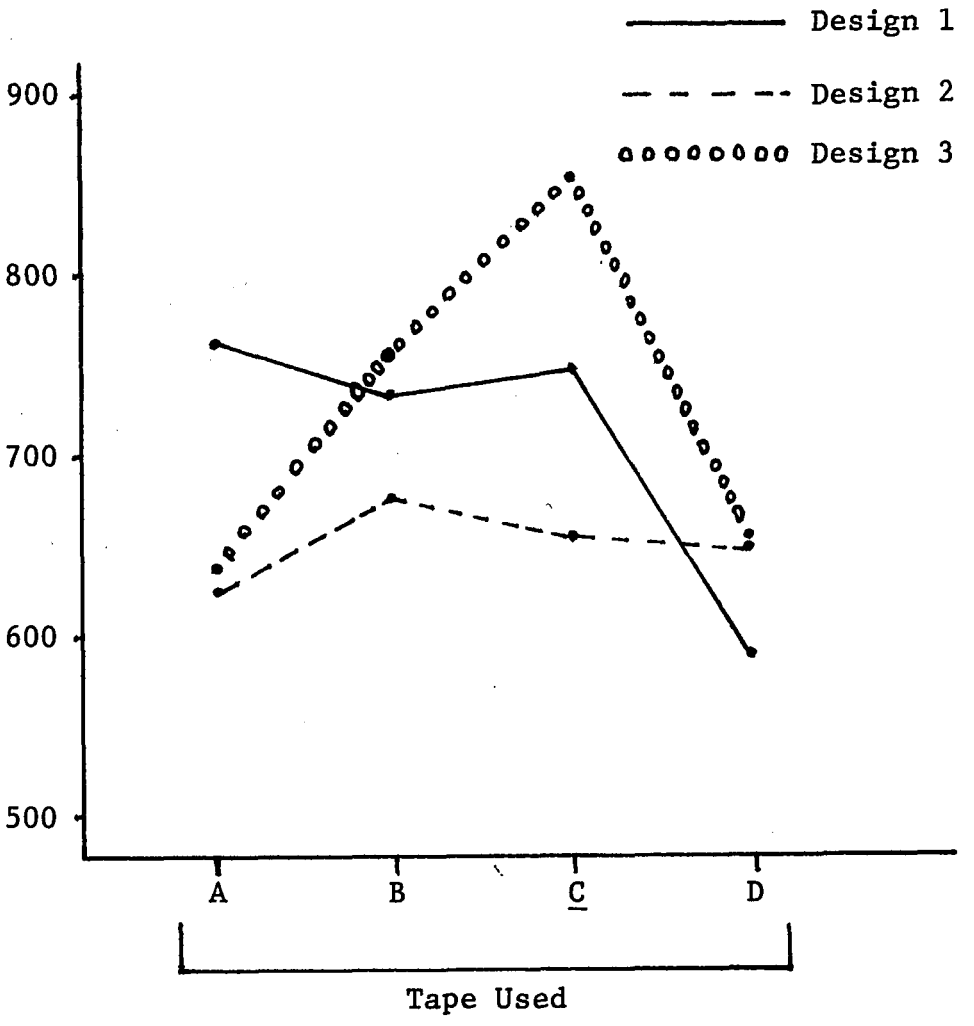


Figure 10. Profiles across the four groups for each design.

It is recommended therefore, that the analysis of the data from filler sentences become as routine a procedure as the inclusion of filler sentences in psycholinguistic experiments. This recommendation requires that the filler analyses have at least as much power as the experimental analyses. If, under those circumstances, the filler analyses support the assumption of equivalent groups, the experimental design is more internally valid than if the data from the filler sentences had not demonstrated the equivalence of groups. The researcher who has found such equivalence is able to make a more confident statement about the liklihood of a causal relationship between the independent and dependent variables.

If the filler analyses demonstrate that the groups are not a priori equivalent, then the situation is more complicated. Information about group differences which are independent of experimental treatment effects can be used as a covariate to adjust experimental mean squares. Keppel (1973) outlines the logic and assumptions of an analysis of covariance, and also presents the formulae necessary for computation.[32] Conceptually, the information about the pre-experimental relationship among the groups is partialed out of the experimental analysis, and the analysis then proceeds "as if" the groups were initially equivalent.

The pattern of results in the filler analysis is compared with the pattern of results in the experimental

---

[32]See especially p. 480, where Keppel outlines the various situations that may warrant the use of an analysis of covariance.

analysis. If the order of the groups is the same in the two analyses, then the experimental results are spuriously high due to differences between the groups that are not related to the independent variable. The focal experiment is not internally valid. Without knowledge about group performances on the filler sentences the group differences would be erroneously ascribed to differences in treatment. If the differences demonstrated by the filler analyses are of the same magnitude and direction as the experimental effects, then a Type I interpretation error could occur; that is, differences would be attributed to treatment effects when, in fact, the treatment variables are not responsible for the differentiation of the groups.

If, on the other hand, the order of the groups in the filler analyses is different from the order of the groups in the experimental analyses, then there is reason to suspect that the results of the experimental analysis are an understatement of the effects of the treatment variables. Suppose an analysis of the filler sentences shows that the reaction times of Group A are significantly greater than those of Group B, and an analysis of the critical sentences shows that the reaction times of Group B are significantly greater than those of Group A. In this case, the implication is that the treatment associated with Group B is even more powerful than the experimental analysis suggests since Group B had to overcome its initial deficit relative to Group A.

Within the Fisher model of analysis of variance, a nonsignificant result may be due to either no effect or to

a lack of power. It is similarly possible that a significant result may be due to either a true effect, or to an excessively powerful test of significance. It is well known that, other things being equal, the addition of subjects increases the liklihood of obtaining significant results. This means that given enough subjects--enough power--any difference between groups will be significant.[33] This is one reason why significance alone cannot be the sole attribute of good research. Part of a researcher's task is to construct a design that has enough power to detect real differences, but not so much power that trivial differences become "significant." A design that includes an analysis of the data from filler sentences can, as we have seen, address this problem. Consider the following three statistical occurrences:

1. The filler analysis is nonsignificant and the experimental analysis is significant--as in the present study. In this case if the power of the two designs is comparable--as it is in the present study--then the researcher has empirical evidence that the treatment differences are not artifacts that result from an excessively powerful design.
2. Both analyses are nonsignificant. There are, in this case, two possibilities: either the independent variable

---

[33]Meehl (1967) presents a striking illustration: he and Lykken have collected data on 45 "miscellaneous variables" from 55,000 subjects. Analysis of that data indicates that 91% of all possible pairwise comparisons are judged to be statistically different by tests of significance.

had no effect or the test lacked power. If adding more subjects (or more materials for an F_2 analysis) results in significant effects for the treatment variables but does not change the results of the nonsignificant filler analysis, then the conclusion must be that the added subjects increased the power of the design so that true effects could be detected.

3. If adding more subjects results in significant filler and experimental analyses (or if both analyses were originally significant), then the following possibilities exist: either the design is too powerful, and the effects are statistical artifacts, or the added power was necessary to demonstrate the treatment effects. In the latter case, however, the treatment effects are confounded by the initial nonequivalence of the subject groups. A comparison of the pattern of results in the filler and experimental analyses is necessary. As was pointed out earlier, an analysis of covariance may be appropriate.

In sum, data from filler sentences represents useful information available to the psycholinguistic researcher. Importantly the filler analyses can shed light on aspects of the internal validity of the experimental design. Moreover, conclusions about treatment effects are less ambiguous when they are integrated with a consideration of the results of the filler analyses. Any study that includes filler sentences in the presentation of the experiment, but not in the analysis of the results, is therefore incomplete.

Experimental Psycholinguistics: Exploration (Design 3)

The major difference between the fillers of design 2 and design 3 is that in design 3 subject groups are not treated alike: although each subject group in design 3 hears the same sentences, the probe words for each group differ. There is no attempt in design 3 to investigate the equivalence of subject groups. Rather, this design, like design 1 (critical sentences), assumes the equivalence of subject groups. Thus, an interpretation of the effects of the various probe types is in part determined by the results of design 2. Since the a priori equivalence of the groups, or at least the lack of a context-related nonequivalence of the groups can be inferred from the analyses of design 2, the differences between the groups (Figure 8) can be linked to the Type of Probe that the groups received. Remember, however, that design 3 represents a preliminary investigation into the sensitivity of the probe task to nongrammatical variables. Thus inferences from this design must be quite tentative.

It was hypothesized that, if the words within a clause are stored and retrieved as a function of their temporal order, then the mean reaction time for the ADJACENT-IN group ought to be greater than the mean reaction time for the IN group. Results from Newman-Keuls analyses did not confirm this hypothesis. The equality of the two groups suggests that information about the contents of a clause is stored in

a wholistic rather than elemental fashion. More recent information is therefore not, in this case, more easily accessible. Although clause order does seem to be a perceptually salient aspect of sentence interpretation, the order of words within the clause is not necessarily perceptually salient. Of course, it should be noted that probed words were selected solely on the basis of their temporal positions in the sentence. Words 18, 12 and 6 syllables from the end of the sentence were probed. These positions were chosen so that subjects would not anticipate that the probe would be from any particular part of the sentence. A consequence of this selection criterion is that probe words for both the IN and ADJACENT-IN conditions are from various syntactic categories: nouns, pronouns, adjectives, adverbs, verbs (main and auxiliary), conjunctions, and prepositions. In addition, the structural relationship between the IN and ADJACENT-IN probes varies considerably: sometimes the pair is from a common constituent, sometimes not; nouns and modifying adjectives are probed as are subjects and verbs, prepositional phrases, etc. It might be that when adjacent words are part of the same construction (i.e., a noun phrase like "young man"), they are stored and retrieved as a unit. Perhaps adjacent words that are in separate constituent structures (i.e., in a subject-verb relationship, the subject is in a noun phrase, and the verb is in a verb phrase) are stored and retrieved differently. There are too few probe pairs in each category to examine the question with the present data. It may well be

that specific elements within a clause are stored and retrieved on a temporal basis, but the conditions for such a process need to be specified. It is not a general rule that all elements within a clause retain their temporal markings during a retrieval task.

The nonsignificance of the difference between the IN and ADJACENT-IN groups has an interesting implication. When the syntactic relationship between the probe pairs is variable, there is no appreciable difference in the reaction times to adjacent words. In the initial investigation of the clause-boundary effect, Caplan (1972) specifically compared subjects' reaction times to adjacent words that are separated by a clause boundary. In this case, reliable differences occurred. By combining the information from these two findings, the reaction-time differences in the clause-boundary study can be more confidently interpreted. They are more likely attributable to the syntactic variable that assigns the probe words to separate clauses rather than to the differences in the temporal order of the two words. Caplan later avoided possible temporal-order confounding by using sentence pairs with different syntactic structures, but common lexical items, and then probing the same word in each pair member. Results from design 3 (that the IN group does not differ significantly from the ADJACENT-IN group) suggest that this added control may not be necessary. Any complex sentence can be used to further investigate the clause-boundary effect since the temporal order alone is not a complete explanation of reaction-time

differences to adjacent words. Researchers perhaps then need not restrict themselves to sentences that have different syntactic constructions but common lexical items. By using a wider variety of sentences, the artificiality of the experimental material will be reduced, and the external validity of the experiment will be increased.

In design 3, groups 1 and 4 (the IN and ADJACENT-IN groups) had to respond as quickly as possible to probe words from the sentence. Recall that for design 2 (filler sentences) all groups performed a similar task--responding as quickly as possible to probe words that were usually from sentence (14 out of 20 times). Consider the performances of groups 1 and 4 on the related tasks of designs 2 and 3 as replications (see Tables 11 and 16 and Figure 10):

	Design 2	Design 3
Group 1	622.23	640.80
Group 4	644.81	662.67

The comparable mean reaction times of the groups across similar tasks suggests again that groups treated alike do not differ--that goes for different groups receiving the same treatment or for the same group receiving multiple-like treatments.

With the exception of the lack of a difference between the IN and ADJACENT-IN groups, the remainder of the predicted relationships between groups were supported by the Neuman-Keuls analyses. The mean reaction time for the OUT group was

significantly larger than the mean reaction times for either the IN or ADJACENT-IN groups. In order to determine that a probe word is not from the sentence just presented, a subject must scan the whole sentence. For the IN and ADJACENT-IN conditions, when the probe was from the last clause, only the last clause (the most accessible one to immediate memory) need be scanned. In both conditions, probes were from the last clause in half the sentences. Since subjects in the IN and ADJACENT-IN groups had to scan both clauses only half the time, and subjects in the OUT group had to scan both clauses all the time, it is obvious that the mean reaction time for the OUT group would be longer. In addition, there is additional evidence for the suggestion that the behavior of the subjects in the OUT group is less variable than the behavior of the IN and ADJACENT-IN groups; the standard deviation for the OUT groups is the lowest of any of the groups (Table 21).

There is a situation in which OUT responses might be faster than IN responses. An incidental finding by Lang (1974) suggests that the shortest reaction times result from out probes that are content words and clearly outlandish with respect to the meaning of the sentence. For example:

24. Yesterday you gave me a very beautiful present and today you gave me another one. CARBOHYDRATE

Presumably, listeners use a heuristic in which the probe is compared to a more wholistic representation of the semantic content of the sentence. The limits of how unrelated a probe must be to the semantic content of the sentence before the heuristic is invoked has yet to be investigated.

The present study however does suggest that, if the probe word rhymes with a word in the sentence, the phonetic similarity complicates the task and increases the reaction time. The mean reaction time for the RHYME group is significantly larger than each of the other three group means. Limber and Lang (1974) found that probes which rhyme with a word in the sentence substantially increase the complexity of the probe task for individuals learning English as a second language. The present study is aimed at investigating this "confusion effect" in native speakers. The magnitude of confusion that language learners experience can then be interpreted relative to the conclusions of this study. Moreoever, if a complete profile can be developed for native speakers which indicates not only the independent effects of various types of probes, but also relational effects among those probe types, then it is quite likely that the probe paradigm can provide a useful index of the listening comprehension skills of individuals learning English as a second language. The probe task allows one to investigate the automatic, unconscious aspect of language use, the aspect typically least investigated by most standardized tests of English as a foreign language.

In sum, the conclusions derived from the analyses of Probe Types (design 3) are of a preliminary and speculative nature. The results suggest that the probe paradigm can be a useful tool for investigating the structure of the internal representation of a sentence. In addition, comparative

analyses of data from probe experiments with native speakers and those learning English might be profitably used to assess the proficiency of language learners.

Generality in Psycholinguistic Research: A Problem Reconsidered

Students in "Introduction to Research Methodology" encounter the distinction between "descriptive" and "inferential" statistics. The former are numbers (means, standard deviations, etc.) that summarize a particular set of data. The latter are tests of significance that reveal the relationship between the particular set of data and potential sets from hypothetical replications. Inferential statistics are aimed at quantifying the probability of replication.

In empirical research, hypothesized relationships are examined only on small samples. If conclusions from such research were limited to the particular elements sampled, little would be gained. Rather, presumably representative samples are studied, and with the use of inferential statistics, conclusions are offered about the populations from which the samples are drawn. Tests of significance involve, therefore, a mechanical approach to the problem of induction (see Introduction, pp. 35-41).

Scientific Research: "Soaked in Theory"

Interesting and useful research is thus necessarily inferential. Yet is is a mistake to equate the inferential

research with inferential statistics. Statistics or tests of significance are prescriptive procedures. They provide the researcher with a set of rules and regulations which, taken together, represent a formal, standardized approach to induction. Induction, however, may be far too problematic to submit to a ritualized procedure. As Hume argued, inductive inferences are always based on insufficient evidence; one can never be sure that what is inferred in any way reflects what exists. This is especially true of theoretical inferences; there is no procedure which insures that theoretical generalizations are coherent and consistent with one's results. Moreover, it is a mistake to assume that theorizing occurs (and should occur only) after the data is collected and statistical significance established, though many psychologists still believe, apparently, that they put on the theoretician's hat only when they discuss their results (see e.g. Bakan, 1966; Rozeboom, 1960; Greenwald, 1975). In fact, it appears that the entire scientific enterprise, from first observations to final proofreading is, as Popper has said, " . . . soaked in theory" (1976, p. 132). Philosophers of science who agree on little else still conclude unanimously that theorizing comes first and last in science (e.g. Kuhn, 1970; Polany, 1958; Popper, 1962). Inductive inferences, long the preoccupation of experimental psychologists, and now of psycholinguists, always occur in the service of a theory.

Induction is both the pride and the problem of science. The aim of research is to infer beyond the informa-

tion given; but it is difficult to establish for those inferences "beyond." Fisher (1955, 1960) and Clark (1973, 1976) have suggested that statistical criteria--tests of significance--can solve this problem (see Introduction, pp. 37-40). Clark has expanded on the Fisher model of analysis of variance by reminding researchers that they are simultaneously sampling from both subjects and materials populations. He believes that, unless appropriate statistical procedures are used (ex: quasi-F or min F' analyses), the inferences "beyond" have no statistical support and so are "unreliable" and "can lead to serious error" (1973, p. 335). Wike and Church (1976) have rejected Clark's statistical refinements and proposed instead that the limits of induction should be established empirically by means of replications (see Introduction, pp. 41-45).

Clark (1973, 1976) and Wike and Church (1976) suggest different solutions to the problem of induction. Yet implicit in the two approaches is a belief that the problem of induction can be solved by prescriptive rules (be they statistical or empirical). Some specific procedure is outlined, and adherence to this procedure is a necessary prerequisite for reliable inductive inferences. Induction becomes a consequence of some mechanical sequence. Either an F statistic or a count of successful replications is held to be the necessary and sufficient criterion for induction. Both approaches neglect the role of theory and thus each is inadequate.

Kaplan (1964) has captured the irony and futility of psychologists' excessive dependence on procedure:

> . . . in contemporary behavioral science the attitude toward experimentation is in danger of becoming a kind of ritualism as though the laying on of hands can itself effect a cure of diseased ideas. As with all rituals, the emphasis passes from content to form, from substantive questions to procedural ones, and virtue comes to be localized in the proper performance of fixed act sequences. (p. 146)

Using the present study as an example, it will be demonstrated that results from statistical and empirical manipulations are of interest only when those maneuvers are embedded within an explicit theory.

The Inadequacy of a Purely Statistical Solution: The Min F'. Clark (1973) has suggested that ". . . almost everyone . . . is committing the language-as-fixed-effect-fallacy" (p. 355)--that is, that conclusions from research are likely to be based on spuriously high F-ratios that result from testing effects with an inappropriate error term (one that is too small). By "doing the right statistics" (p. 347), researchers are safe from Type I error epidemic (incorrect rejection of Ho) that can result from the fixed-effect-fallacy. In studies that have not adhered to these procedures, Clark implies that a reanalysis will probably show that significant results are limited to the linguistic sample investigated. Both Clark (1973) and his critics (Wike and Church, 1976) have described quasi-F and min F' analyses (the right statistics) as conservative tests. To Clark, this is a desirable trait, necessary in order to avoid Type I errors. To his

critics, conservative tests are to be avoided since their usage can lead to an epidemic of Type II errors (incorrect failure to reject Ho).

The Use of Power. Although the min F' is a conservative test because it mathematically underestimates the value of the quasi-F ratio, it is not necessarily an inherently stringent test. In and of itself, the min F' neither encourages nor discourages Type I or Type II epidemics. The min F' test is a set of sequential procedures that blindly operate on strings of numbers. It is part of the research process to deliberately coordinate a specific design with a particular statistical test. In order to effectively make use of any statistical test, researchers must therefore be aware of the concept of power, the probability of correctly rejecting the null hypothesis. A proper consideration of power involves fitting the strength of an hypothesized effect and the sensitivity of the statistical tests used to measure that effect. In order to set up an experimental design and choose an appropriate test for the analysis, researchers must thus be knowledgeable about the theoretical assumptions that underlie statistical testing, and have a theoretical and/or empirical basis for hypothesizing a specific magnitude of effect. The popular but misconceived notion of a very stringent test is really a case of an inappropriate power match between design and analysis, in particular, the design may lack power.

The statistical remedies suggested by Clark (quasi-$\underline{F}$ and min $\underline{F}'$ tests) appear to be overly stringent tests because the analysis often does not fit the design (see Introduction, pp. 41-45). When past studies are reanalyzed, originally significant effects become nonsignificant. It is not necessarily that the effects do not generalize across language materials, but rather that the design was not explicitly constructed to test that possibility. Both quasi-$\underline{F}$ and min $\underline{F}'$ analyses are new to most researchers. If they blindly apply the new computational formulae to traditional designs, they should hardly be surprised by failures to confirm experimental hypotheses. The addition of a second random factor to the design (as is required by quasi-$\underline{F}$ and min $\underline{F}'$ tests) results in a substantial reduction of power. Researchers not familiar with the new procedures will not be able to take this into account when designing their experiments. In Clark's (1973) description of his newly proposed min $\underline{F}'$ test (which is a combination of the subjects analysis and the materials analysis), he implies that the sensitivity of the test is purely a statistical matter: "an experimental design is only as sensitive as the less sensitive of the two subdesigns it contains" (p. 349). But that cannot be true. It is not statistical tests, per se, that are sensitive or insensitive, but the minds of researchers. Those minds contain theories (implicit or explicit), and it is on the basis of those theories that researchers select their design and analysis.

Clark is quite correct in pointing out that psycholinguistic researchers have failed to consider the implications of sampling from a population of materials as well as from a population of subjects. Indeed, it is obviously true that research is carried out on a small sample of linguistic materials, but inferences are applied to a much larger set. The legitimacy of the inference however, does not rest ultimately on the outcome of a test of significance. Results of a test of significance are of interest only if the power of the test is large enough to detect true effects (differences that are theoretically meaningful), but not so large that trivial differences are judged significant.[34] Achieving this match between an experimental design and a test of significance requires more than the blind adoption of a new statistical procedure. It requires an understanding of the properties and how they interact with the properties of experimental design. If a new procedure which treats two factors as random leads to a loss of power, the researcher must know this and then consider aspects of experimental design that will increase the power of the test.

In the present study an attempt was made to equate the relative power of the subjects analysis (F_1) and the materials analysis (F_2). As a preliminary step, the number of subjects was equated with the number of materials. However,

---

[34]Again, the example described in Meehl (1967) and in footnote 33 clearly illustrates the importance of distinguishing between differences that are only statistical judgments and differences that are both theoretical and statistical judgments.

the subjects analyses were between subject designs, while the materials analyses were repeated measures designs. Thus the materials analysis is more powerful than the subjects analyses. But this is a satisfactory situation since Clark (1973) has implied that it is often the materials analysis that is the weaker of the two and so provides the upper bound for the min F'. Note, however, that for the critical sentences, the F values from the subjects analyses are smaller than the F's from the materials analyses (see Tables 3 and 6). In future research, if the subjects design is between subjects and the materials design is a repeated measures, perhaps the number of materials could be less than the number of subjects.

In this situation, where the statistical tests of significance are relatively new and unfamiliar, the methodological role of the filler sentences is emphasized. The numbers of subjects and sentences are the result of an educated guess about the strength of hypothesized effects when measured by subjects and materials analyses. Thus, it is quite important that the design in which the fillers are analyzed be as powerful as that of the experimental analysis. Otherwise, subject and material numbers could be excessively increased to the point where Type I errors would be a strong possibility.

The combination of significant experimental results (critical sentences--design 1 and probe types--design 3) and nonsignificant filler results (design 2) suggests that, in the present study, the power of the tests of significance was large enough to detect meaningful differences, but not so

large as to be sensitive to trivial differences. Moreover, the significant results of this study make it clear that, when attention is paid to both the subject and material factors of a design, an analysis that considers both factors as random need not be thought of as excessively stringent (see Table 10).

Defining a Population. Clark (1973, 1976) has presented statistical procedures for analyzing a design that contains two random variables. In addition to the traditional subjects variable, Clark recommends that researchers consider the materials variable to be random. Otherwise, in his view, psycholinguistic researchers commit the fixed-effect-fallacy:

> Investigators . . . almost never provide statistical evidence that their findings generalize beyond the specific sample of language materials choices. Nevertheless, these same investigators do not hesitate to conclude that their findings are true for language in general. (1973, p. 335)

The implication is that if the "right statistics" are used, then, and only then, may researchers unhesitatingly conclude that "their findings are true for language in general."

Clark is, as we have seen, concerned only with the statistical basis for making generalizations. If the min F' is large enough, one may generalize findings; if the min F' is not large enough, one may not so generalize. Yet, even if researchers present statistical evidence that they generalize beyond their sample, one may still ask: to what population should the generalization be applied? Clark's prescription for generalizing to "language in general" is surely not correct. The experimental materials are never selected from

"language in general." A useful definition of linguistic populations must come from a theoretical conception of the variables being investigated.

In the present study, all experimental analyses resulted in significant min $\underline{F}'$ values (see Table 9). The effects of the Probe Position and Type of Clause variables are presumed to be reliable across changes in both subject and material populations. But what exactly are those populations? Before the implications of the analyses were discussed, salient characteristics of the populations were made explicit (see Discussion, pp. 116-117).

Defining a population is not an unambiguous task. Many alternative descriptions can correctly characterize any set of language materials. Decisions regarding which characteristics are or are not salient, are contingent upon the theoretical psycholinguistic knowledge of the researcher. For example, it is not obvious from an inspection of the critical sentences (see Appendix A) that syntactically complex lexical items were used. Within the framework of a psycholinguistic theory of sentence processing, however, it is meaningful to group lexical items according to their syntactic and perceptual complexity (see Introduction, pp. 25-28, on lexical analysis). Thus, the deliberate use of syntactically complex lexical items implies that the attribute of perceptual complexity is a necessary part of the description of the population. It may well be that the results generalize to populations of sentences with less complex lexical items, but the

consideration of that possibility is beyond the scope of this present study.

Although the statistical significance of a study cannot change with time, its theoretical significance can. Theoretical advances often lead to changes in the way the materials are described and therefore to changes in the population to which the results are generalized. For example, Lang (1974) does not mention that all the sentences used in her probe study contain verbs that are syntactically and perceptually simple. It is possible that findings from that study are applicable only to sentences from a population of perceptually simple verbs. Conversely, future research may suggest that the variables of Probe Position and Type of Clause operate independently of the types of lexical items contained in the sentence. If so, results from the present study will have added generality.

In sum, the results of appropriate tests of significance can provide useful information about the probability of replication, but only under specified conditions: if a different sample of subjects is drawn from the same population; if a different sample of materials is drawn from from the same population. It is, however, the researcher--not the test of significance--who defines the population. That definition, moreover, is based on a theoretical understanding of the variables under investigation.

The Inadequacy of a Purely Empirical Solution: The Replication. Levy (1969) has described the myth of a perfect

replication. Changes from study to study are inevitable. But, the question is, are these changes to be interpreted as "nuisance" variables, evidence of the limitations of behavioral research, or are some at least to be interpreted as intentional variations that are the consequences of explicit attention to theory. It is only the results from experiments that incorporate some of this latter form of variation that can be incorporated into an argument describing the generality of a phenomenon.

In the present study, interest in the variables of clause order and type of clause is based on a theory of sentence interpretation that explicitly incorporates both structural and perceptual aspects of language (see Introduction, pp. 2-34). Within this conceptual framework, the results of a number of studies have been compared (see Introduction, pp. 46-51, and Discussion, pp. 116-124). It is from those studies that the variables of interest in the present study have emerged. The salient characteristics of the populations from which the samples were derived have been described and suggestions have been made for future research which can both replicate and extend the present findings (see Discussion, pp. 116-139).

A Theoretical Approach to Induction: Problems, Not Solutions

In sum, inductions from samples of subjects and materials to broad populations are not made more plausible by

rigid adherence to ritualized "fixed act sequences." Clark's major contribution, therefore, is not that he has told researchers which error term to use in order to properly analyze experimental results. Rather, his critique of psycholinguistic experimentation has usefully directed the attention of researchers to the need to explicitly consider the materials factor. It is also clear, however, that researchers need to go beyond Clark to a realization that a statistical treatment of materials presupposes an explicit theoretical treatment of the materials factor. Similarly, the suggestions of Wike and Church are useful not because their replicatory prescriptions can, in part or in sum, help a researcher to increase the generalizability of experimental results. They instead have simply reminded researchers of a viable nonstatistical method for collecting evidence of generality. But this method also presupposes an explicit theoretical treatment of the relevant variables (including the materials factor).

Researchers must present an argument for the generality of their results. Statistical and empirical evidence may be used in the argument, but the evidence is not the argument, per se. Rather, such arguments derive their force from the theoretical sense they make.

Summary of Results and Suggestions for Future Research

Clausal, lexical and surface-structure analyses are an integral part of any theory of sentence interpretation. Using the probe paradigm, design 1 of the present study examined the effects of two variables related to the description of adverbial and main clauses in complex sentences: the temporal order of the clause (whether a probe was from the first or second clause of the sentence) and the type of clause (whether a probe was from the subordinate or main clause of the sentence). In addition, although only complex sentences with adverbial clauses were considered, the lexical items of each sentence readily allow for the later construction of complement-clause versions of each sentence. Thus, the sentences with adverbial clauses frequently contained lexical markers (main verbs, head nouns, modal auxilaries, or adjectives) that could dominate a complement construction. All initial subordinate clauses contained surface-structure cues that marked the embedding.

It was found that although the main effects for both variables were significant (the reaction time to first-clause probes was longer than the reaction time to second-clause probes; and the reaction time to probes from subordinate clauses was longer than the reaction time to probes from main clauses), a significant interaction qualified those results. In particular, the effect of clause order was significant

only for main clauses, and the effect of type of clause was significant only for clauses in the second position.

These results suggested a storage and retrieval model for sentential material in which clausal information that is less accessible (either because of its first position temporal order, or because of its subordinate construction, or both) is held in a kind of undifferentiated storage. Material from a clause which is both recent and main is placed in one category; material from all other clauses is placed in another less accessible category. This may be the way the temporal order and type-of-clause variables interact in all situations; or, it may be that this pattern is restricted to complex sentences with adverbial clauses that contain syntactically complex lexical markers. Future research can explore the effects of syntactically simple lexical items and other clausal constructions (relative and complement).

The use of filler sentences in the present experiment increased the external validity of the experiment. (External validity refers to the extent that the sample or laboratory situation is an accurate copy of the population or "real world" to which one wishes to generalize the results.) With the inclusion of filler sentences, the array of sentences presented to subjects was not the repetitive structural array that would have resulted from a presentation of just experimental sentences. In addition, when filler sentences were combined with experimental sentences, the heterogeneity of the complete sentence set disguised the intent of the study and

therefore decreased the liklihood that subjects would form uniform and correct predictions about what they should do on succeeding sentences.

For the filler sentences (design 2), all subjects heard the same sentences and the same probe words. The filler sentences were thus a variable of "no treatment" across the four subject groups studied. For the filler analyses, the predictions were for nonsignificant results. In order to infer theoretical support, rather than just a lack of power, from those null results, two other distinct but comparably powerful designs were examined (designs 1 and 3). Significant results were predicted and found for those two designs. Thus, the analyses of the data from the filler sentences was interpreted as support for the hypothesis of the initial equality of the groups; or, by an alternate interpretation, the analyses supported the hypothesis that the different contexts that the filler sentences were presented in (the unique combination of treatments from designs 1 and 3) did not affect the performance of subject groups on an identical task. Both of these interpretations added a measure of internal validity to the study. (Internal validity refers to the extent that a researcher can eliminate competing hypotheses about the reason for the differences among the experimental groups.) The differences between the experimental groups in designs 1 and 3 were more confidently attributed to the manipulated variables rather than to any pre-existing differences between

the groups or to any set differences that might have resulted from the different probe types that each group received.

In design 3, four different probe words were used for each sentence. It was found that probes not from the sentence (both OUT's and RHYME's) evoked longer reaction times than probes from the sentence (both IN's and ADJACENT-IN's). One plausible interpretation of these results is that, in order to determine that a particular word was not from that sentence, the entire sentence must be scanned, while the task of determining the inclusion of a probe word from the sentence is completed as soon as that word is scanned. Thus, scanning an entire sentence should take longer than scanning only a portion of it.

Probes that rhyme with a word from the sentence, evoked longer reaction times than probes unrelated to any word in the sentence. The phonetic similarity between the probe word and a word in the sentence adds a measure of confusion to the task. This confusion effect has also been found in students learning English as a second language (Limber & Lang, 1975). Future research can consider a direct comparison of the magnitudes of the confusion effect for native speakers and second-language learners. It may well be that the magnitude of such an effect can be used as an index of listening comprehension skills.

A comparison of the reaction times to probe words that are adjacent to each other in the sentence was included in this research as a preliminary step toward investigating

within-clause variables involved in sentence processing. The finding that the IN and ADJACENT-IN group reaction times did not significantly differ from one another suggests that the variable of temporal order of words within a clause should be studied more subtly. In particular, the perceptual variable of word order should be examined under separate structural conditions such as adjective-noun, subject-verb, verb-object, etc.

In this exploratory research, the effects of only four different probe words were compared (IN, OUT, RHYME, ADJACENT-IN). Future experimentation with other probe types can shed light on additional aspects of sentence interpretation. The possible confusion effect of synonomous probes (probes that are semantically related to a word in the sentence) has yet to be systematically investigated with either native speakers or second-language learners. A study of the effect of outlandish probes (probes that are obviously unrelated to the message of the sentence) could lead to the formulation of a heuristic that listeners use: when the probe word is clearly unrelated to the semantic content of the sentence, don't scan the whole sentence, just guess that the word was not from the sentence. Again the performance of native speakers and second-language learners could be compared.

For all three designs, both subjects and language materials were conceptualized and analyzed as random factors. Results from a subjects analysis (F_1) and a materials analysis

($\underline{F}_2$) were combined to produce min $\underline{F}'$ statistics for each effect. In all cases, effects that were significant in one analysis were significant in all analyses; effects that were nonsignificant in one analysis were nonsignificant in all analyses. Results supported the view that, if the power of the underlying $\underline{F}_1$ and $\underline{F}_2$ analyses are comparable, then the min $\underline{F}$; is not an overly stringent test of experimental hypotheses; although it is a mathematically conservative test (that is, it underestimates the quasi-$\underline{F}$ statistic).

The results from the three designs are thus reliable over the subject and material populations sampled. The statistical procedures associated with tests of significance do not, however, specify the populations sampled. It is the researcher, drawing on the theoretical framework that surrounds the study, that defines the populations that the findings generalize to. Similarly, empirical generalization can be achieved when the results of research findings are integrated with an explicit theoretical framework to produce suggestions for future research. Both the populations investigated, and directions for future research, have been outlined.

In summary, the probe paradigm is a versatile research tool. It can be used to investigate clausal, lexical, and surface-structure analyses in sentence interpretation. Further, salient within-clause structures and strategies can be explored. Finally, explicit attention to methodological aspects of design and analysis will increase the confidence

that a researcher has in the internal and external validity of the study.

BIBLIOGRAPHY

Bakan, D. The test of significance in psychological research. Psychological Bulletin, 1966, 66, 423-437.

Barber, T. X. Invalid arguements, postmortem analyses, and the experimenter bias effect. Journal of Consulting and Clinical Psychology, 1969, 33, 11-14.

Barber, T. X., Calverley, D. S., Forgione, A., McPeake, J. D., Chaves, J. F., & Bowen, B. Five attempts to replicate the experimenter bias effect. Journal of Consulting and Clinical Psychology, 1969, 33, 1-6.

Bever, T. G. A survey of some recent work in psycholinguistics. In W. J. Plath (Ed.), Specification and utilization of a transformational grammar: Scientific report number three. Yorktown Heights, N.Y.: Thomas J. Watson Research Center, International Business Machines Corp., 1968.

Bever, T. G. The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), Cognition and the development of language. New York: Wiley, 1970.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Caplan, D. Acoustic store and perception of sentences. Unpublished paper, M.I.T., 1970.

Caplan, D. Clause boundaries and recognition latencies for words in sentences. Perception and Psychophysics, 1972, 12, 73-76.

Chomsky, N. Syntactic structures. The Hague: Mouton, 1957.

Chomsky, N. Aspects of the theory of syntax. Cambridge, Mass.: MIT Press, 1965.

Chomsky, N. Language and mind (2nd ed.). New York: Harcourt, Brace & Jovanovich, 1972.

Clark, H. H. The language-as-fixed-effect-fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 335-359.

Clark, H. H. Reply to Wike and Church. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 257-266.

Clark, H. H., & Clark, E. V. Semantic distinctions and memory for complex sentences. Quarterly Journal of Experimental Psychology, 1968, 20, 129-138.

Coleman, E. B. Generalizing to a language population. Psychological Reports, 1964, 14, 219-226.

Evans, R. I. B. F. Skinner: The man and his ideas. New York: Dutton, 1968.

Fisher, R. A. Statistical methods and scientific induction. Journal of the Royal Statistical Society, (B), 1955, 17, 69-78.

Fisher, R. A. The design of experiments (7th ed.). New York: Hafner, 1960.

Fodor, J. A. The language of thought. New York: Crowell, 1975.

Fodor, J. A., & Bever, T. G. The psychological reality of linguistic segments. Journal of Verbal Learning and Verbal Behavior, 1965, 4, 414-420.

Fodor, J. A., Bever, T. G., & Garrett, M. F. The psychology of language. New York: McGraw-Hill, 1974.

Fodor, J. A., & Garrett, M. F. Some reflections on competence and performance. In J. Lyons & R. J. Wales (Eds.), Psycholinguistic papers. Edinburgh: University of Edinburgh Press, 1966, 135-179.

Fodor, J. A., & Garrett, M. F. Some syntactic determinates of sentential complexity. Perception and Psychophysics, 1967, 2, 289-296.

Fodor, J. A., Garrett, M. F., & Bever, T. G. Some syntactic determinants of sentential complexity, II: Verb structure. Perception and Psychophysics, 1968, 3, 453-461.

Forster, K. I., & Dickinson, R. G. More on the language-as-fixed-effect-fallacy: Monte Carlo estimates of error rates for F1, F2, f' and min F'. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 136-142.

Fromkin, V., & Rodman, R. An introduction to language. New York: Holt, Rinehart & Winston, 1974.

Garrett, M. F., Bever, T. G., & Fodor, J. A. The active use of grammar in speech perception. Perception and Psychophysics, 1966, 1, 30-32.

Greenwald, A. G. Consequences of prejudice against the Null hypothesis. Psychological Bulletin, 1975, 82, 1-20.

Hume, D. An inquiry concerning human understanding (L. A. Selby-Biggs, Ed). In R. M. Hutchins (Ed.), Great books of the western world (Vol. 35). Chicago: Encyclopedia Brittanica, 1952, 449-509. (1748)

James, S. The principles of psychology (2 vols.). New York: Holt, 1890.

Kaplan, A. The conduct of inquiry. San Francisco: Chandler, 1964.

Katz, J. J. The realm of meaning. In G. A. Miller (Ed.), Communication, language and meaning: Psychological perspectives. New York: Basic Books, 1973, 36-48.

Keppel, G. Design and analysis: A researcher's handbook. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Kornfeld, J. R. The influence of clause structure on the perceptual analysis of sentences. Unpublished Ph.D. thesis, M.I.T., 1974.

Kuhn, T. S. The structure of scientific revolutions (2nd. ed.). Chicago: University of Chicago Press, 1970.

Ladefoged, P., & Broadbent, D. E. Perception of sequence in auditory events. Quarterly Journal of Experimental Psychology, 1960, 12, 162-170.

Lang, J. M. Investigations in sentence processing. Unpublished master's thesis, University of New Hampshire, 1974.

Levy, L. H. Reflections on replications and the experimenter bias effect. Journal of Consulting and Clinical Psychology, 1969, 33(1), 15-17.

Limber, J. Syntax and sentence interpretation. In E. C. T. Walker & R. J. Wales (Eds.), New approaches to language mechanisms. Amsterdam: North Holland, 1976, 151-181.

Limber, J., & Lang, J. M. Sentence processing in monolinguals and bilinguals using a probe latency paradigm. Paper presented at the Conference on Language and Learning, Queens College, New York, June 1975.

Lykken, D. T. Statistical significance in psychological research. Psychological Bulletin, 1968, 70, 151-159.

McGuigan, F. J. The experimenter: A neglected stimulus object. Psychological Bulletin, 1963, 60, 421-428.

Meehl, P. E. Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science, 1967, 34(2), 103-115.

Miller, G. A., Galanter, E., & Pribram, K. H. Plans and the structure of behavior. New York: Holt, 1960.

Miller, G. A., & McKean, K. A chronometric study of some relations between sentences. Quarterly Journal of Experimental Psychology, 1964, 16, 297-308.

Neale, J. M., & Liebert, R. M. Science and behavior: An introduction to research methods. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Neisser, U. Cognitive psychology. New York: Appleton-Century-Crofts, 1967.

Orne, M. T. On the social psychology of the psychology experiment: With particular reference to demand characteristics and their implications. American Psychologist, 1962, 17, 776-783.

Polanyi, M. Personal knowledge: Toward a post-critical philosophy. New York: Harper, 1958.

Popper, K. R. Conjectures and refutations: The growth of scientific knowledge. New York: Basic, 1962.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Rosenthal, R. On not so replicated experiments and not so null results. Journal of Consulting and Clinical Psychology, 1969, 33, 7-10.

Rozeboom, W. W. The fallacy of the null-hypothesis significance test. Psychological Bulletin, 1960, 57, 416-428.

Sidman, M. Tactics of scientific research. New York: Basic Books, 1960.

Slobin, D. I. Grammatical transformations and sentence comprehension in childhood and adulthood. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 219-227.

Watson, J. B. Behaviorism. New York: Norton, 1930.

Weksel, W., & Bever, T. G. Harvard Cognitive Studies Progress Report, 1966.

Wike, E. L., & Church, J. D. Comments on Clark's "The language-as-fixed-effect-fallacy." Journal of Verbal Learning and Verbal Behavior, 1976, 15, 249-255.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

Wundt, W. Principles of physiological psychology (5th ed., Vol. 1). (E. B. Titchener, trans.) New York: MacMillan, 1904. (1874)

APPENDIX A

List of Critical Sentences

These sentences are revisions of adverbials used by Kornfeld (1974). All sentences are 24 syllables long and the probe word is always 12 syllables from the end.

1. LAND

   a. When the greedy rancher began purchasing land, farming plots became quite scarce in Suffolk county.

   b. The greedy rancher tried to purchase farms when land was becoming more and more scarce in Suffolk county.

   c. The greedy old rancher tried to purchase more land when farms were becoming scarce in Suffolk county.

   d. When the greedy old rancher began purchasing farms, land quickly became quite scarce in Suffolk county.

2. CAR

   a. Since a highly-skilled mechanic worked on the car, the engine was back in good running condition.

   b. The mechanic worked on the engine since the car was not in a very good running condition.

   c. The mechanic worked for a long time on the car since the engine was in very bad condition.

   d. Since the mechanic worked on the engine, the car is again in very good running condition.

3. TIME

   a. If you can prepare the meal ahead of time you can relax a little while before dinner.

   b. You could prepare the meal ahead, if you want time for leisurely conversation before dinner.

   c. You surely could prepare the meal ahead of time if you want to take a short rest before dinner.

d. If you can cook the meal ahead, you will have time for leisurely conversation before dinner.

4. JUDGE

a. Because our lawyer was rude in front of the judge, we naturally became extremely angry.

b. Our lawyer should have been less rude because the judge of the court was becoming more and more angry.

c. Our lawyer should have been much less rude with the judge because the trial was important to us.

d. Because our lawyer was so rude in court, the judge of the case naturally became quite angry.

5. FARMERS

a. Ever since feed costs increased greatly for farmers, there have been grave problems in the economy.

b. Feed costs have been increasing ever since farmers have had many other economic problems.

c. Feed costs have been increasing greatly for farmers ever since the recent economic problems.

d. Ever since feed costs have increased greatly, farmers have had many other economic problems.

6. PLANTS

a. Ever since we used special insect spray on plants, the leaves have been safe from attacks of small aphids.

b. We've used insect spray on leaves ever since the plants were attacked by several kinds of tiny aphids.

c. We have used special kinds of insect spray on plants ever since the leaves were attacked by small aphids.

d. Ever since we used insect spray on leaves, the plants have been safe from most attacks of tiny aphids.

7. HIGHWAYS

a. Because the repairs had started on the parkways, heavy traffic was jammed up during the rush hour.

b. Some of the repairs were stopped because the parkways jammed up with heavy traffic during the rush hour.

c. Some of the road repairs were stopped on the parkways because traffic was jammed up during the rush hour.

d. Because the road repairs had started, the parkways jammed up with heavy traffic during the rush hour.

8. CHIEF

a. Because the rookie gave false reports to the chief, the officers on the police force were upset.

b. The rookie checked in right away because the chief of the police force was very upset with him.

c. The new officer had to report to the chief, because the entire force was upset with him.

d. Because the officer gave false reports, the chief of the police force was very upset.

9. SQUAD

a. After the sergeant gave the order to the squad, the men were ready for the enemy soldiers.

b. The sergeant gave several orders after the squad was surprised by attacking enemy soldiers.

c. The sergeant gave several quick orders to the squad, after the men were surprised by the enemy.

d. After the master sergeant gave orders, the squad was ready for all of the enemy soldiers.

10. SINGER

a. After the band quarreled often with the singer, she was finally fired by the director.

b. The band had many quarrels, after the singer was finally fired by the director himself.

c. The band had many quarrels about the singer, after she was fired by the director himself.

d. After the band had many quarrels, the singer was finally fired by the director himself.

11. CLASS

a. Because there was always so much noise from the class, the teacher was in trouble with the principal.

b. There should really be much less noise, because the class could get into trouble with the new principal.

c. There should really be a lot less noise from the class, because the teacher could get into much trouble.

d. Because there was so much noise from the room, the class was in serious trouble with the principal.

12. GROUNDS

a. Because that lazy caretaker ignored the grounds, the yard soon turned into an overgrowth of weeds.

b. That caretaker ignored the yard because the grounds were too big to keep free of weeds and overgrowth.

c. That lazy caretaker ignored most of the grounds, because the yard was too big to keep free of weeds.

d. Because the caretaker ignored the yard, the grounds soon turned into an overgrowth of ugly weeds.

13. BOOKS

a. Now that efficient clerks have audited the books, the company's accounts are kept right up to date.

b. The clerk audits efficiently now that the books of the company are all kept right up to date.

c. The new clerk can efficiently audit the books, now that the accounts are all kept right up to date.

d. Now that the clerk audited the accounts, the books of the company are all kept right up to date.

14. HOUSE

a. If you'll put new screens on the windows of the house, the rooms will be free of insects all summer long.

b. You might put new screens on the windows if the house is usually full of most kinds of insects.

c. You might put new screens on the windows of the house if the rooms are usually full of insects.

d. If you'll put up new screens on the windows, the house will surely stay free of insects all summer long.

15. WALLS

a. Because the nurse hung some bright pictures on the walls, the rooms looked a lot more cheerful to the patients.

b. The nurse hung up some bright pictures, because the walls of the rooms would look cheerful to many patients.

c. The nurse hung up many bright pictures on the walls because the rooms would then look a lot more cheerful.

d. Because the nurse hung up some bright pictures, the walls of the rooms looked more cheerful to many patients.

16. COFFEE

a. Although retail prices are quite high for coffee, tea continues to sell at a good bargain price.

b. Prices are very high for tea, although coffee is very much a bargain by comparison.

c. Many retail prices are quite high for coffee, although tea continues to be quite a bargain.

d. Although retail prices are high for tea, coffee still continues to sell at a good bargain price.

17. SHOWERS

a. Before swimmers in the gym had used the showers, there was always a large supply of hot water.

b. All the swimmers used the gym before the showers no longer had any hot water supply left.

c. All of the swimmers in the gym used the showers before the hot water supply was depleted.

d. Before many swimmers used the gym, the showers had an ample enough supply of hot water.

18. COEDS

a. Though the dean answered the demands of the coeds, the dorm committee will hardly be satisfied.

b. The dean answered all the demands, though the coeds of the dorm council will hardly be satisfied.

c. The dean should respond to demands of the coeds, even though the dorm will hardly be satisfied.

d. Even though the dean answered the demands, the coeds of the dorm council will hardly be satisfied.

19. MATCH

a. After your partner was confident of the match you both won spots in the national tournament.

b. Your partner was more calm after you won the match easily in the national squash tournament.

c. Your tennis partner was confident of the match after you both won the national tournament.

d. After your partner was more calm, you won the match easily at the national squash tournament.

20. PLOT

a. Even before many knew details of the plot, Watergate was brought to the public's attention.

b. Politicians knew the details before the plot of Watergate was brought to public attention.

c. A few politicians knew details of the plot even before the Watergate was made public.

d. Before many knew all the details, the plot of Watergate was brought to public attention.

APPENDIX B

List of Filler Sentences

Design 2

All sentences are 24 syllables long. All subjects heard the same sentences and the same probes.

Probes 18 syllables from the end:

S/M	1.	Whenever I don't eat a good balanced meal, I feel much more tired and run-down than usual.	EAT
	2.	Even though you are not an Italian, you really seem to enjoy eating Italian cuisine.	NOT
M/S	3.	Let's go out to dinner tonight because the afternoon meeting usually gets over late.	DINNER
	4.	It was pleasant to sit by the glowing fireplace while we talked about the events of the past year.	SIT
coordinate	5.	This morning we got up later than than usual but we still managed to get to class on time.	UP
	6.	The local radio station is planning a fund-raising marathon and they need volunteers.	RADIO
	7.	Their new furniture was delivered and they were surprised by how beautiful the room now looked.	WAS

Probes 6 syllables from the end:

S/M	8.	Whenever I feel the least bit tired during the day, I take a nap for an hour or so.	NAP

	9.	Because the spring term ends in three weeks, some students are already at work on their final projects.	WORK
M/S	10.	The young man always polished his brown shoes with such care because he wanted them to last a long time.	WANTED
	11.	Jim brewed a cup of hot cinnamon tea for her because he noticed she was was feeling a bit cold.	SHE
coordinate	12.	Today we will go to the dress designer and we will select the style for the special long dress.	STYLE
	13.	I have read your beautiful letter many times, and each time it brings me more pleasure than before.	ME
	14.	That couple will be married soon, and so they are both very busy with wedding details right now.	WITH

Probes not from the sentence:

S/M	15.	Whenever we are in a hurry, we often seem to forget about the most important things.	CUTE
M/S	16.	The children continued to enjoy their new record even though they had played the songs many times.	TIRED
	18.	Grandma really likes to get letters from both of us, and we enjoy writing to her very much.	SHOE
	19.	We went for long walks over the weekend, but we also spent time relaxing by the fireplace.	BLUE
	20.	The students complained about the difficulty of the test, but the instructor would not listen.	SNOW

APPENDIX C

List of Filler Sentences

Design 3

All sentences are 24 syllables long. Each sentence has four different probes. Each subject heard only one probe type.

Probes 18 syllables from the end:

S/M 21. Because the tall, young man was so strikingly handsome, he looked terrific in all kinds of clothing.

MAN YOUNG PAN JOB

22. While we are on our walk tonight, let's see if any of the neighborhood houses have their lights on.

WALK OUR TALK TOAST

M/S 23. You should sit down and rest immediately whenever you begin to feel the least bit dizzy.

REST AND GUEST CHEESE

24. They both did a good job of cleaning the house today because they had invited guests for dinner.

JOB GOOD ROB SMILE

coordinate 25. We are planning our trip to Colorado this summer, but first we have to save up the money.

TRIP OUR SLIP HAIR

26. Martha likes to walk three miles every day, but sometimes when it's snowing or raining she does not walk.

THREE WALK FREE SOX

Probes 12 syllables from the end:

S/M 27. Because of the exam next Monday, I will not be able to go on the ski trip this weekend.

NOT WILL POT PLANT

28. Often, when you are teaching in the morning, I like to go over to the library and read.

I MORNING SKY JUICE

M/S 29. I am ready to do a lot of walking this winter because I bought a new pair of warm boots.

THIS WALKING KISS FACE

30. I felt very quiet and relaxed when we were listening to the classical music last night.

WERE WE FUR RABBIT

coordinate 31. Six months ago Jim decided to let his beard grow and now he needs to trim it just once a week.

BEARD HIS WEIRD LETTER

32. They were both very hungry when they got home, but soon after supper they were full and satisfied.

BUT HOME CUT GOLD

33. The instructor spent hours preparing for class, and all his lectures were thoughtful and organized.

CLASS FOR GLASS CHAIN

34. We will fly to New England in a month but right now we have to work out the travel arrangements.

RIGHT BUT HEIGHT REVIEW

Probes 6 syllables from the end:

S/M 35. Whenever you begin reading one book, it always reminds you of things you've read in other books.

THINGS OF STRINGS SNACK

36. After the big fire in the neighborhood, people checked to see if they had enough insurance.

THEY IF CLAY CLOCK

M/S 37. The young man was becoming quite impatient because his tiny cut did not seem to be healing.

DID CUT HID WINK

38. Let's go to the library for a few hours because we both have much work to do by tomorrow.

WORK MUCH JERK TEACH

coordinate 39. Your weekend visit was very enjoyable, and the letter you wrote later was quite thoughtful.

WROTE YOU QUOTE DRESS

40. Tonight we should write a letter to Ronda and we should send it air mail special delivery.

MAIL AIR SAIL HOUND

APPENDIX D

Random Order of Stimulus Sentences

C1 through C20 represent the critical sentences.

F1 through F20 represent the filler sentences of design 2.

F21 through F40 represent the filler sentences of design 3.

Tape Position	Sentence Number	Tape Position	Sentence Number
1.	F39	31.	F 7
2.	F 2	32.	C 9
3.	F30	33.	F 9
4.	F26	34.	C12
5.	C13	35.	F22
6.	F 6	36.	F 3
7.	F13	37.	F23
8.	F25	38.	C 3
9.	C10	39.	C17
10.	F18	40.	F11
11.	C19	41.	F38
12.	F20	42.	F40
13.	C 4	43.	F16
14.	C18	44.	C14
15.	F28	45.	F29
16.	C11	46.	F 4
17.	F 8	47.	F12
18.	F24	48.	F31
19.	C 2	49.	C20
20.	F20	50.	C 7
21.	F17	51.	F 1
22.	C 8	52.	F36
23.	F21	53.	F34
24.	F33	54.	C 6
25.	F14	55.	F37
26.	F 5	56.	C15
27.	F19	57.	F15
28.	C 5	58.	C 1
29.	F27	59.	C16
30.	F35	60.	F10

APPENDIX E

Instructions

This is an experiment in auditory perception. You will hear a series of test sentences. After each sentence, you will hear a word which may or may not have been in the previous sentence. Your task is to push the lever to the position marked "IN" if the word was from the previous sentence or to the position marked "OUT" if the word was not. It is important to respond as quickly as possible. If you think that you made a mistake, you may correct it before going on to the next sentence.

In order to help you, a short tone has been placed just after the sentence. You must judge if the word that follows the tone was in the previous sentence. After you have responded, you will hear another tone that will signal the beginning of the next sentence.

The word that follows the sentence may be a word from the sentence (in which case you would respond IN) or it may be a word totally unrelated to any words from the sentence (in which case you would respond OUT), or it may be a word that rhymes with a word in the sentence (again, however, in these cases you would respond OUT).

Turn over the index card in front of you and look at the three example sentences:

1. In spite of the college population, Allendale is a small town. COLLEGE
2. This is a good area to shop for antiques. PENCIL
3. The new rug makes our living room much cozier. HUG

In the first sentence, the test word COLLEGE is from the sentence. Therefore, the correct response would be IN. PENCIL is not in the second sentence, so the correct response would be OUT. For the third sentence, although HUG rhymes with rug, the word hug is not in the sentence, so OUT would be the correct response. Please turn over the index card.

Two of the three sentences will not be presented as if they were the test sentences. Please push the lever to the appropriate IN or OUT position as quickly as you can after you hear the test word. If you are ready, the two example sentences will be presented.

In spite of the college population, Allendale is a small town. COLLEGE

This is a good area to shop for antiques. PENCIL

In addition to responding IN or OUT immediately after you hear the test word, two or three times during the experiment you will also be asked to paraphrase the sentence that you just heard.

We are now ready to begin the experiment. Remember that your primary task is to press the lever to the IN or OUT position as quickly as possible after you hear the test word.

Do you have any questions?